UNIVERSITÀ DEGLI STUDI DELL'AQUILA

DIPARTIMENTO DI INGEGNERIA E SCIENZE DELL'INFORMAZIONE E MATEMATICA

Dottorato di Ricerca in Ingegneria e Scienze dell'Informazione

Curriculum Systems Engineering, Telecommunications and HW/SW Platforms

XXXIV ciclo

Titolo della tesi

An Optimal Transport Perspective on the Approximation of Mixture Densities

SSD ING-INF/04

Dottorando

Alessandro D'Ortenzio

Coordinatore del corso	Tutor
Prof. Vittorio Cortellessa	Prof. Costanzo Manes

a.a. 2020/2021

Contents

List of Figures

List of Algorithms

Acronyms

ABTI Average Barycentral Triangular Identity. 101, 102, 108, 148, 224, 231

ADF Average Dissimilarity Function. 99, 117, 118, 139

AIC Akaike Information Criterion. 56, 194, 200, 207, 208

ATC Accuracy Threshold Criterion. 6–8, 153–157, 159–164, 186–192, 194, 198, 204

AWGN Additive White Gaussian Noise. 72

BC Bhattacharyya Coefficient. 92

BD Bhattacharyya Distance. 90

BIC Bayesian Information Criterion. 8, 56, 194, 196, 198, 200–203, 207, 208

BSDA Best Single Density Approximation. 98, 99, 102, 103, 167, 171, 180, 206

BSGA Best Single Gaussian Approximation. 112, 114, 116, 119

CA Constant Acceleration. 71

CC Chernoff α-coefficient. 92

CD$'_\alpha$ Chernoff α-Divergence of the I° kind. 93

CD$''_\alpha$ Chernoff α-Divergence of the II° kind. 95

cdf cumulative distribution function. 31, 32, 98, 115

CIPs Cross Information Potentials. 85

CKF Cubature Kalman Filter. 72

CM Correlation Measure. 90, 132

CSD Cauchy-Schwarz Divergence. 90, 169

CT Coordinated Turn. 71

CTD Composite Transportation Dissimilarity. 136, 143, 165–167, 206, 209

CV Constant Velocity. 71

DA Data Association. 76

DoF degrees of freedom. 41

EF Exponential Family. 37, 107

EKF Extended Kalman Filter. 72

EM Expectation-Maximization. 4, 8, 63, 64, 66, 67, 107, 122, 132, 133, 163, 194–204

EMD Earth Mover's Distance. 134

EOT Extended Object Tracking. 181

ETT Extended Target Tracking. 182

FKLD Forward Kullback-Leibler Divergence. 83

FPI Fixed-Point Iteration. 100, 103, 112, 113, 115, 116, 118, 119, 121, 139, 169, 173, 178, 208

GGIW Gamma Gaussian Inverse-Wishart. 4, 7, 8, 52, 181–192, 207, 208, 235

GM Gaussian Mixture. 7, 24, 64, 65, 171, 172, 175–178

GMM Gaussian Mixture Model. 24

GMPHD Gaussian Mixture PHD. 79

GMs Gaussian Mixtures. 9, 65, 177, 179, 180

GS Gaussian Sum. 79

H2 Square Hellinger Distance. 94

HEM Hierarchical Expectation Maximization. 132, 133

i.i.d. independent and identically distributed. 41, 42

ISD Integral Squared Difference. 88

ISE Integral Squared Error. 88

JR2D Jensen-Renyi Quadratic Divergence. 91

KDE Kernel Density Estimation. 195

KF Kalman Filter. 72, 73

KLA Kullback-Leibler Average. 111

KLD Kullback-Leibler Divergence. 82

KLI Kullback-Leibler Mean Information. 82

L2 Square L_2 norm. 88

LB Likeness-Based. 85, 87, 99, 116–119, 122, 208

MAP Maximum a Posteriori. 50, 51

MC Monte Carlo. 177, 178

MD2 Square Mahalanobis' Distance. 43

MI Mutual Information. 132

ML Maximum Likelihood. 44, 46, 48, 49, 51, 62, 107

MLE Maximum Likelihood Estimation. 42, 50, 51, 107, 145, 167, 194

MM Majorization-Minimization. 64, 66, 165

MoDs Mixture of Densities. 58

MoGs Mixture of Gaussians. 24

MOT Multi-Object Tracking. 79

MR Mixture Reduction. 140, 146, 163, 169, 178, 182, 209

MRP Mixture Reduction Problem. 25, 26, 79, 84, 96, 118, 123, 133, 135, 136, 139, 140, 143, 167–170, 184, 205, 206, 208, 209

ND Neyman χ^2 Divergence. 94

NISE Normalized Integral Squared Error. 89

NLL Negative Log-Likelihood. 42–44, 46, 48

NN Nearest Neighbour. 79

NR Newton-Rhapson. 46, 48

OTP Optimal Transport Problem. 134–136, 138–140, 143, 146, 206

OTT Optimal Transport Theory. 25, 26, 134, 145, 165, 166, 169, 184, 206–208

PD Pearson χ^2 Divergence. 94

PDA Probabilistic Data Association. 79

pdf probability density function. 32, 38–41, 50, 58, 107, 108, 215

PF Particle Filter. 72

PHD Probability Hypothesis Density. 193

PMBM Poisson Multi-Bernoulli Mixture. 79

pmf probability mass function. 31, 32, 77

PPP Poisson Point Process. 75, 76

RFS Random Finite Set. 181, 192

RFSs Randon Finite Sets. 24, 79

RKLD Reverse Kullback-Leibler Divergence. 83

RTL Relative Transportation Loss. 7, 146–148, 150–153, 155, 156, 187–190

rv random variable. 31

SKLD Symmetrized Kullback-Leibler Divergence. 84

SMC Slope Magnitude Criterion. 6–9, 153–157, 159–164, 171, 186, 187, 189, 191, 192, 194, 196, 198, 200, 203, 204, 209

SPD Symmetric Positive-Definite. 160

TSL Total Squared Loss. 89

UKF Unscented Kalman Filter. 72

W2 Squared 2-Wasserstein. 85, 114

List of Symbols

$\mathbb{N}$	Set of natural numbers
$\mathbb{N}_0$	Set of natural numbers including zero
$\mathbb{R}$	Set of scalar real numbers
$\mathbb{R}_+$	Set of positive-only real numbers
$\mathbb{R}_+^n$	Set of positive-valued vectors in $\mathbb{R}^n$
$\mathbb{R}^d$	Set of d-dimensional real vectors
$\mathbb{R}^{d\times d}$	Set of d-dimensional real square matrices
$\mathbb{R}^{m\times n}$	Set of $m \times n$ real rectangular matrices
$P(x)$	Cumulative distribution function
$\Pr(\cdot)$	Probability of an event
I_n	n-dimensional identity matrix
$\|\cdot\|$	Euclidean norm
$\langle\cdot,\cdot\rangle_F$	Frobenius norm
$\mathrm{vec}(\cdot)$	Stack operator of the columns of a matrix into a vector
$[1\!:\!n]$	The interval of integers from 1 to n
$\perp\!\!\!\perp$	Independent symbol
$\mathcal{N}^d$	The set of nondegenerate Gaussian pdfs on $\mathbb{R}^d$
$\mathcal{Q}$	Generic family of distributions (pdfs)
$\mathcal{Q}_{\mathrm{mix}}$	Mixtures of distributions (pdfs) with components in $\mathcal{Q}$
$\mathcal{Q}_{\mathrm{mix}}^{(n)}$	Mixtures of distributions (pdfs) with components in $\mathcal{Q}$ with exactly n components
$\mathcal{U}_{[a,b]}$	1-dimensional Uniform distribution in the interval $[a,b]$
$\mathcal{U}_{[a,b]}^d$	d-dimensional Uniform distribution in the hyper-cube of side $[a,b]$

$\nu(x)$	(or simply $\nu \in \mathcal{N}^d$) Generic Gaussian density (compact form)
$\nu(x\|\mu, \Sigma)$	Gaussian density of mean μ and covariance Σ
ν_i	Compact form for a Gaussian density of mean μ_i and covariance Σ_i
$\boldsymbol{\nu}$	A vector of Gaussians, whose components are denoted ν_i, $i = [1\!:\!n]$
$\mathcal{G}$	The set of Gamma pdfs
$\gamma(\chi\|\kappa, \omega)$	Gamma density of shape parameter κ and scale parameter ω
$\gamma(\chi)$	Generic Gamma density (compact form)
$\gamma_i(\chi)$	Generic Gamma density of shape parameter κ_i and scale parameter θ_i
$\mathcal{IW}^d$	the set of d-dimensional inverse-Wishart pdfs
$\varphi(\mathcal{Y}\|V, v)$	Inverse-Wishart density of matrix parameter V and scalar degrees of freedom v.
$\varphi(\mathcal{Y})$	Generic inverse-Wishart density (compact form)
$\varphi_i(\mathcal{Y})$	Generic inverse-Wishart density of matrix parameter V_i and degrees of freedom v_i
$\Gamma(\cdot)$	Euler Gamma function
$\Gamma_d(\cdot)$	Multivariate Euler Gamma function
$\|\cdot\|$	Determinant of a matrix
S^d_{++}	Open convex cone of symmetric positive definite (SPD) matrices
$\mathbb{1}_n$	Vectors of ones in $\mathbb{R}^n$
$\boldsymbol{w}$	d-dimensional weight vector
Δ^{n-1}	Standard simplex, subset of $\mathbb{R}^n_+$
$\bar{\ell}(\boldsymbol{x}\|\theta)$	Negative log-likelihood of the data set $\boldsymbol{x}$ given the parameter(s) θ
η	Natural parameters in the exponential family
$t(x)$	Sufficient, or natural, statistic in the exponential family parameterization
$a(\eta)$	Log-partition function in the exponential family parameterization
$h(x)$	Base measure in the exponential family parameterization

$\psi_0(\cdot)$	Digamma function (a.k.a. polygamma function of order zero)
$\psi_1(\cdot)$	Trigamma function (a.k.a. polygamma function of order one)
$\psi_d(\cdot)$	Polygamma function
$\bar{\Phi}_D(\cdot)$	Function computing the barycenter of a set of weighted components

Acknowledgement

First and foremost, a thank goes to my supervisor, Prof. Costanzo Manes, for all the support provided during my doctorate. I would like to thank my family, in particular my sister Simona, my friends Luis, Ilaria, Alessio, and Valentina, for being there when I needed the most. I would like to thank Dr. Francesco Capelli and Dr. Silvia Mammarella, for the excellent support provided during the last year. I would also like to thank the radio transmission La Zanzara, which represented a source of endless laughs during my darkest hour. A thank goes to Prof. Umut Orguner, who has proven to be a very kind and inspiring person. A thank goes to all the FUSION community, for introducing me to the topics of filtering and target tracking. Last, but not least, I would like to thank all the people who took part and inspired me during the last four years.

The Office - s07e19, 14:45.

Foreword

My academic experience has been really puzzling. I have always ended up doing stuff which was rather too difficult for me, or which I would have gladly avoided to do. At first, I wanted to study computer engineering, because I was not very skilled in maths, whereas I was able to code decently; nonetheless, I have ended up earning a Bachelor's degree in control systems engineering. During my Master's degree, I have got really passionate about robotics and filtering theory, but I was held back by the lack of sound probability theory knowledge, which led me to avoid any topics involving statistics. Somehow, the fate has led me to start a Ph.D. during which I have been able to study systems identification, statistics, probability theory and optimization, which represented, basically, the antithesis of my original study plans. Nonetheless, during the past four years, concepts which would have appeared absurd to me in the beginning, started to get more clear over time, until, at some point, I have started to grasp the whole picture. Target tracking and unsupervised learning became my core topics after attending an enlightening course held by Prof. Lennart Svensson from the Chalmers university; there, I have encountered for the first time the so called Mixture Reduction problem. Battling with such a problem pushed me towards interesting topics as Optimal Transport theory which, jointly with probability, statistics and information theory, represented the foundations for the results reported in this dissertation. In the end, the Ph.D. has been my most rewarding educational experience, both in terms of personal growth and acquired knowledge. Right before the "finish line", during the 25th International Conference on Information Fusion, held in Linköping, Sweden, I have received the "Tammy L. Blair Best Student Paper Award" for the paper "A model selection criterion for the mixture reduction problem based on the Kullback-Leibler divergence", written in collaboration with my supervisor Costanzo Manes, and the distinguished Professor Umut Orguner. Such an award has meant a lot to me, since it has proven, in a small part, that my efforts have led to something useful for the related scientific community. This thesis is the result of a work carried out at the University of L'Aquila (L'Aquila, Italy) during the years 2018-2022.

Abstract

Many real world problems deal with data generated from sources of which laws are usually unknown. In order to make predictions or to infer the corresponding behavior, suitable probabilistic models are often sought to characterize such systems. Nonetheless, when the underlying nature is complex, simple models might result to be unsatisfactory, and it is necessary to resort to more descriptive forms; an efficient, yet powerful, alternative is to consider mixture models which, by combining simpler elements, allow to characterize particularly complex behaviors. However, when the number of mixture parameters becomes significantly large, such models can become computationally intractable and approximations have to be introduced. The goal of this dissertation is to address the mixture reduction problem in an *Occam's razor* perspective, that is by finding good trade-offs between representation complexity and accuracy. The proposed methodology is general, but, for the goals of this work, the focus will be posed on target tracking in clutter problems, where the optimal Bayesian recursion leads to an unbounded increase in the number of mixture components, making corresponding algorithms intractable, and on clustering problems, especially in high dimensional settings, where finding a suitable number of representatives is a non-trivial task. Given a mixture model with many components, the corresponding reduction problem consists in finding another mixture, possessing a significantly less components, which is similar, in some sense, to the original one. The first step to address this problem is to have a measure of how dissimilar two mixtures are; in the literature many dissimilarity measures have been proposed, each of which possessing its own features, but a discussion regarding the corresponding peculiarities has been rarely addressed. Moreover, many of those dissimilarities are analytically intractable when applied to mixtures, leading to the reasonable approach of resorting either to tractable approximations, or to employ an ensemble of measures in the same reduction pipeline in or-

der to ease the computations. In this work, a mixture reduction scheme is defined consistent when the same dissimilarity measure is employed for each of the steps involved in the process; in this regard, by exploiting the Optimal Transport theory, it is possible to formulate a consistent reduction framework which is also capable, in a specific case, to deal with the corresponding model selection problem, that is to provide automatically a suitable number of components for the reduced order model; choosing the order of the reduced model can be an impactful choice, since overly simple approximations can lead to a large bias of the representations, and overly complex alternatives can be computationally burdensome. In this dissertation, the optimal transport theory will serve as a systematic approach to solve all-round the mixture reduction problem, by providing all the tools necessary to reduce and refine a given mixture model.

Chapter 1

Introduction

When dealing with real world problems, it is common to exploit mathematical models in order to make predictions or to characterize an underlying, potentially unknown, behavior. Many physical phenomena, though, can not be described deterministically exhaustively, since they can exhibit random dynamics; in that case, probabilistic models can be considered. Nowadays, fields like robotics [2], machine learning [3] or target tracking [4] rely massively on characterizing statistically the processes underlying the data or the system dynamics, given that such representations can be a good trade-off between complexity and versatility. A key concept for random phenomena is that of *uncertainty* which, among several interpretations, can be seen as the information one lacks when describing a system.

Probability theory [5] provides a solid framework to quantify and manipulate the uncertainty, since it offers tools and models to characterize and understand random events. Complementary, statistics [6] allow to extrapolate useful information from such models and to make inference or predictions regarding a given system. Among the mathematical tools offered by statistics, the Bayesian framework has gained a lot of attention during the past decades, since many estimation or decision problems can be posed as a recursive or batch application of the *Bayes' rule* [3]. In contexts like target tracking or robotics, for instance, the state of a dynamical system is often modelled as Gaussian, i.e. one does not know the position of the object exactly, but it is possible to provide a guess, in terms of probability, about its average value and potential deviations from it; in such settings, the dynamical state can be estimated by means of Bayes filters (e.g. Kalman filter [7]). Nonetheless, the Gaussian assumption can be limiting, in the sense that, often, the state of a system can not be accurately described by a single Gaussian component: the real (unknown) distribution can either be multimodal or possess a particularly complex geometry.

A common approach to describe multimodal or complex distributions is to consider a convex (weighted) sum of parametric distributions in the same class, e.g., belonging to the *exponential family*, of which the Gaussian distribution is part of. Such a representation is denoted as *mixture of densities*, and it provides an efficient and versatile tool to deal with uncertainty in a broad range of problems. Mixture models are probability density function (pdf) themselves, hence it is required for the weights to add up to one; if such a constraint is removed, then the resulting representation is called *intensity*, that is an *unnormalized* sum of pdfs. Intensities are gaining a lot of attention in the context of target tracking, since state-of-the-art filters, based on Randon Finite Sets (RFSs) [8], exploit such uncertainty description to characterize both the number of objects and their dynamical state.

A commonly used mixture model in fields like machine learning and/or statistical analysis [3, 9, 10], target tracking [11–13] or image retrieval and registration [14, 15], is that of Gaussian Mixture Model (GMM), also known as Mixture of Gaussians (MoGs) or GM, given that it results to be a particularly efficient, yet accurate, uncertainty description tool for many practical scenarios. For instance, if one considers tasks like localizing a robot in a map or tracking several targets in the presence of clutter[1], and where an optimal Bayesian estimation approach is considered together with the Gaussian assumption for the system state, then GMs arise naturally due to the presence of phenomena like *data association uncertainty*[2], multiple models[3], multimodality of the state or the measurement noise and so on. In such settings, though, the number of components in a GM is subject to an exponential growth, leading the representation complexity to increase unbounded over time: the uncertainty description becomes intractable after few steps.

[1] In target tracking, we denote by clutter the *unwanted* information observed by the sensors due, for instance, to the presence of cross-talking, interference, reflections and so on.

[2] When several measurements are received at a given time step, and no information about which target (if any) is responsible for their generation is available, then it is necessary to consider all the combinatorics between the present objects and such set of observations when performing a filter update.

[3] When modelling the dynamics of *maneuvering* targets, a common approach is to consider multiple dynamical models which should catch all the possible maneuvers of an object of interest.

In this regard, keeping the uncertainty description tractable is a crucial point, especially in real-time contexts where computational resources are limited and decisions have to be taken in a finite time. If the random dynamical state of a system is represented by a mixture of densities, it would be useful to approximate the potentially intractable pdf with a corresponding *reduced* or simpler form, which is similar, in some sense, to the original representation. Such a problem falls under the name of the Mixture Reduction Problem (MRP).

During the past three decades, the MRP has been tackled by exploiting different approaches, but it can still be considered an open problem from several points of view. The vast majority of the existing solutions address the MRP by trying to minimize one, or multiple, loss functions between an original and a reduced model through a set of subsequent steps: a general structure is to consider a Greedy reduction of the components followed by a refinement phase. Nonetheless, many of those algorithms are often based on heuristics which are not backed by theoretically sound concepts and which, as mentioned, tend to minimize several different loss functions in the same MRP, causing *inconsistency* of the reduction pipeline. With *consistent* (alternatively *coherent* or *congruent*) is denoted a mixture reduction routine where all the steps are aimed at minimizing a single loss function, or dissimilarity measure (D-measure for short), of interest. Nevertheless, each D-measure has its own peculiarities, and the corresponding choice can be impactful from several points of view, like availability of closed forms, preserved features and so on. Another open issue in the MRP is the choice of the reduced model order, that is the number of components the approximated mixture should have in order to provide a good trade-off between accuracy and complexity (Occam's razor); in this regard, an order selection method will be discussed for a particular case of the MRP.

Given the constantly growing employment of mixture models in practical problems, and given the absence of a general method of approaching the MRP, it would be of interest to have a reference framework to address such a problem; in this regard, the goal of this dissertation is to provide some preliminary, yet theoretically sound, contributions in order to tackle the MRP more rigorously and from a general perspective. As result, by exploiting the Optimal Transport Theory (OTT) a general framework is presented which is able to deal systematically with both mixture and intensity models. As it will be discussed further in this work, its key features are:

- *consistency* of the reduction pipeline: all the steps are performed according to a single dissimilarity measure, hence aiming to minimize only a given loss function between mixtures, with the goal of obtaining superior approximations accordingly.

- *efficiency*: for a broad range of dissimilarity measures, the resulting algorithms are suitable for real time applications.

- *versatility*: the framework here presented is not restricted to a specific class of distributions (e.g. Gaussians), but it can be employed for mixtures composed by any kind of distributions and, moreover, it is not restricted to mixtures, since it can be applied directly even to intensities. The only requirement is that the pairwise dissimilarity between components can be evaluated.

- For a specific choice of dissimilarity measure, such a framework provides an embedded model selection criterion capable of halting the reduction during the greedy descent.

The remainder of this dissertation will be organized as follows: in Chapter 2 theoretical fundamentals for the subsequent discussions will be provided, in Chapter 3 the MRP is formulated, a review of the literature is reported and some aspects regarding its solutions are discussed. In Chapter 4 the OTT is introduced and a corresponding reduction framework is proposed. In Chapter 5 some numerical tests are performed to validate the proposed concepts. Conclusions will follow.

Contributing publications and documents to this dissertation

Many of the contents presented in this work were disseminated through the following publications/documents:

- A. D'Ortenzio and C. Manes, "Consistency issues in Gaussian mixture models reduction algorithms," 2021. https://arxiv.org/abs/2104.12586.

- A. D'Ortenzio and C. Manes, "Composite transportation dissimilarity in consistent Gaussian mixture reduction," in IEEE 24th International Conference on Information Fusion (FUSION 2021), Sun City, South Africa, November 2021, pp. 1–8.

- A. D'Ortenzio and C. Manes, "Likeness-based dissimilarity measures for Gaussian mixture reduction and data fusion," in IEEE 24th International Conference on Information Fusion (FUSION 2021), Sun City, South Africa, November 2021, pp. 1–8.

- A. D'Ortenzio, C. Manes, and U. Orguner, "Fixed-point iterations for several dissimilarity measure barycenters in the Gaussian case," 2022. https://arxiv.org/abs/2205.04806.

- A. D'Ortenzio, C. Manes, and U. Orguner, "An optimal transport perspective on gamma Gaussian inverse-Wishart mixture reduction," in IEEE 25th International Conference on Information Fusion (FUSION 2022), Linköping, Sweden, July 2022, pp. 1–8.

- A. D'Ortenzio, C. Manes, and U. Orguner, "A model selection criterion for the mixture reduction problem based on the Kullback-Leibler divergence," in IEEE 25th International Conference on Information Fusion (FUSION 2022), Linköping, Sweden, July 2022, pp. 1–8.

Chapter 2

Fundamental Concepts

In this chapter, the fundamentals concepts will be introduced to provide a self-contained argumentation of the topics which will follow.

2.1 Probability Fundamentals

Let us start with the following sentence:

A fair coin will land heads with 50% of probability.

What does that mean? There exist two main different interpretations of probability, one is the *frequentist* perspective, while the other is the *bayesian* approach. In the former, probabilities represent the frequencies of *events* in the long run, e.g., if a coin is flipped many times, it is expected to land head half of those times. In the latter, the probability is viewed as a way to quantify the *uncertainty* or ignorance about something, and it is related to the concept of available knowledge rather than repeated experiments.

In a Bayesian perspective, the sentence opening this section implies that one is confident that a coin is equally likely to land heads or tails in the next toss [16]. One among the main advantages of the Bayesian framework is that it can be used to model uncertainty related to events which do not repeat often (or maybe repeating just one), or which do not have long term frequencies. In this perspective, based on how probable an event is, one can take the optimal action accordingly. The Bayesian framework, which will be briefly discussed in this chapter, lays beneath many of the modern sciences and applications, like machine learning, robotics, sensor fusion and so on.

Types of uncertainty

The uncertainty about a process can arise for two different reasons. The first is due to the fact that one can be ignorant about the underlying hidden causes or mechanisms generating the observed data. This kind of uncertainty is called *epistemic* (*epistemology*: philosophical term used to describe the study of the knowledge). Another synonym for this kind of uncertainty is *model uncertainty.* The second kind of uncertainty arises from intrinsic variability, which can not be reduced even if more data is collected; this kind is referred to as *aleatoric or data uncertainty.* An example of the difference between the two is the following: consider tossing a fair coin. One might know for sure that it is fair, that is head lands with 50% of probability, so there is no epistemic uncertainty, but one can still not perfectly predict the outcome of a toss.

Probability space

A *probability space* is a triplet $(\Omega, \mathcal{F}, \Pr(\cdot))$, where:

- Ω is a set called *event space*,
- $\mathcal{F}$ is a σ-algebra over Ω, that is a family of subsets of Ω, called *events*, such that:
    1. $\emptyset, \Omega \in \mathcal{F}$,
    2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$, that is, for any event in the σ-algebra, also its complementary belongs to the σ-algebra,
    3. given $\{A_i\}_{i=1}^{\infty}$, $A_i \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.
- $\Pr(\cdot) : \mathcal{F} \to [0, 1]$ is a probability measure, that is a function over sets which provides the probability of an event to take place, such that:
    - $\Pr(\emptyset) = 0$ and $\Pr(\Omega) = 1$
    - given $\{A_i\}_{i=1}^{\infty}$, $A_i \in \mathcal{F}$, such that $A_i \cap A_j = \emptyset$ if $j \neq j$, then

$$\Pr\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \Pr(A_i).$$

Probability of an event

An *event* $A \in \mathcal{F}$ can be thought as some state of the world that either holds or does not hold. Few examples can be "it will rain today", "it rained yesterday" or "the parameter θ is between 1.5 and 2.0". The expression $Pr(A)$ denotes the probability of A being true. It is required that $0 \leq \Pr(A) \leq 1$, with $\Pr(A) = 0$ meaning that the event does not happen, while $\Pr(A) = 1$ denotes that the event is certain. With $\Pr(A^c)$ is denoted the probability of the complement of A, namely A^c, that is the probability of A not happening; this is defined as $\Pr(A^c) = 1 - \Pr(A)$.

Probability of a conjunction of two events

Given two events A and B, the intersection $A \cap B$ is the joint event, and the corresponding *joint probability* of both happening is $\Pr(A \cap B)$. A and B are said to be *independent events* if

$$\Pr(A \cap B) = \Pr(A) \Pr(B). \tag{2.1}$$

Probability of a union of two events

Given two events A and B, the probability of either one of those happening is:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \tag{2.2}$$

If A and B are mutually exclusive, that is they can not happen at the same time, then it holds that:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B). \tag{2.3}$$

Conditional probability of one event given another

Given that an event A has occurred, the *conditional probability* of B happening is defined:

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}. \tag{2.4}$$

Conditional independence of events

Two events A and B are *conditionally independent* given an event C if:

$$\Pr(A \cap B|C) = \Pr(A|C) \Pr(B|C). \tag{2.5}$$

Random variables

Consider some unknown quantity of interest denoted by X, e.g., the outcome of a dice roll or the temperature of a room. If the value of X is unknown and/or could change, then it is possible to call it a random variable (rv). The set of possible values, denoted $\mathcal{X}$, is called *sample space* or *state space*. An *event* is a set of outcomes from a given *event space*.

Discrete random variables

If $\mathcal{X}$ is finite or countably finite, then X is a *discrete random variable*. It is possible to denote the probability that X takes value x by $\Pr(X = x)$.

A probability mass function (pmf), denoted as $p(x) \triangleq \Pr(X = x)$, is a function which computes the probability of events corresponding to a given discrete rv X and which satisfies $0 \leq p(x) \leq 1$ and $\sum_{x \in \mathcal{X}} p(x) = 1$.

Continuous random variables

If X is a real-valued quantity, then it is called *continuous* random variable, that is the set of possible values it can take is no longer finite or countable. Nonetheless, there are a countable number of *intervals* which partition the real line. If the events are mapped to such intervals, then the discussions made for the discrete case still hold. If the size of the intervals can shrink to zero, then it is possible to represent the probability of X of taking on a specific real value; however, in the continuous case, the probability of observing exactly a given real number is zero.

Cumulative distribution function (cdf)

Let us define the events $A = (X \leq a)$, $B = (X \leq b)$ and $C = (a < X \leq b)$, with $a < b$. Then it holds $B = A \cup C$ and, since A and C are mutually exclusive, it holds that:

$$\Pr(B) = \Pr(A) + \Pr(C). \tag{2.6}$$

or, equivalently:

$$\Pr(C) = \Pr(B) - \Pr(A). \tag{2.7}$$

In general, it is possible to define the cumulative distribution function (cdf) of the rv X as:

$$F(x) \triangleq \Pr(X \leq x). \tag{2.8}$$

By exploiting the cdf, it is possible to compute the probability of X being in an interval as:

$$\Pr(a < X \leq b) = F(b) - F(a). \tag{2.9}$$

Cumulative distribution functions are monotonically non-decreasing functions.

Probability density function (pdf)

Once the cdf has been defined, it is possible to define its derivative, called probability density function (pdf), as:

$$p(x) \triangleq \frac{\mathrm{d}}{\mathrm{d}x} F(x). \tag{2.10}$$

This is the equivalent of pmfs for continuous random variables, hence it holds that $\int p(x)\mathrm{d}x = 1$ and $p(x) \geq 0$, $\forall x \in \mathcal{X}$. Given a pdf, it is possible to compute the probability of a continuous rv being in a finite interval as:

$$\Pr(a < X \leq b) = \int_a^b p(x)\mathrm{d}x = F(b) - F(a). \tag{2.11}$$

As the interval size tends to zero, one can write:

$$\Pr(x \leq X \leq x + \mathrm{d}x) \approx p(x)\mathrm{d}x. \tag{2.12}$$

Of course, by collapsing the interval onto a single event (specific real number) the corresponding probability is zero.
Note: in the context of this work, the words distribution or pdf will be used equivalently.

Given the distribution (pdf) of a continuous random variable, one can compute the corresponding cdf as:

$$F(x) = \int_a^x p(x)\mathrm{d}x, \tag{2.13}$$

where a depends on the support of the distribution of interest; for instance, for distributions having the whole real axis as support, one gets $a = -\infty$. In the discrete case, one obtains:

$$F(x) = \sum_{x_i \leq x} p(x_i). \tag{2.14}$$

Marginal distribution

Let us consider two rvs X and Y. One can define the joint distribution as $p(X = x, Y = y)$ for all possible values of X and Y. If X and Y are both discrete, then it is possible to represent the joint distribution by means of a table, otherwise, if continuous, it will result in a continuous function of two variables. From now on, for a matter of compactness, it will be adopted the notation $p(X = x) = p(x)$ or $p(X = x, Y = y) = p(x, y)$ and, when computing integrals for the continuous cases, the integration domain will be assumed to be the whole event space, so the corresponding subscript will be omitted.

Given a discrete joint distribution, it is possible to obtain the *marginal distribution* of an rv as:

$$p(x) = \sum_{y} p(x, y). \tag{2.15}$$

In the continuous case one gets:

$$p(x) = \int p(x, y)\mathrm{d}y. \tag{2.16}$$

By applying the conditional probability rule defined in (2.4), it is possible to write:

$$p(x) = \sum_{y} p(x|y)p(y), \tag{2.17}$$

which for continuous rvs becomes:

$$p(x) = \int p(x|y)p(y)\mathrm{d}y. \tag{2.18}$$

Equations (2.17) and (2.18) are known as the **sum rule** or **rule of total probability**.

It has been used the fact that the *conditional distribution* of an rv can be written as:

$$p(y|x) = \frac{p(x, y)}{p(x)}, \tag{2.19}$$

which, rearranged, becomes:

$$p(y|x)p(x) = p(x, y) = p(x|y)p(y). \tag{2.20}$$

The above equation is called **product rule**, from which it is possible to obtain the following relation:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \tag{2.21}$$

which is known as **Bayes' rule**; note that the denominator of (2.21) can be rewritten, by using the sum rule, as:

$$p(y) = \int p(y|x)p(x)\mathrm{d}x. \tag{2.22}$$

More on this will be discussed in Sec. 2.5.

As done for events, it is possible to define the independence of two rvs X and Y as:

$$p(x, y) = p(x)p(y), \tag{2.23}$$

that is the joint distribution can be factorized as the product of the marginal distributions. Given a set of rvs $X_1 = x_1, ..., X_n = x_n$, they are independent if:

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i). \tag{2.24}$$

It is trivial to generalize the concept of conditional independence; two rvs X and Y are said to be *conditionally independent* given Z if and only if the conditional joint distribution can be factorized as:

$$p(x, y|z) = p(x|z)p(y|z). \tag{2.25}$$

Mean of a random variable

Given a continuous rv $X \sim p(x)$, its *mean*, or *expected value*, often denoted by μ is defined as:

$$\mathbb{E}[X] = \int x \cdot p(x)\mathrm{d}x. \tag{2.26}$$

For a discrete rv X, the mean is defined as:

$$\mathbb{E}[X] = \sum_{x} x \cdot p(x). \tag{2.27}$$

The mean is a linear operator that is it holds:

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b. \tag{2.28}$$

Given a set of n rvs, one can show that the following holds:

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]. \tag{2.29}$$

Moreover, if the rvs are independent, then the expectation of their product can be written as:

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i]. \tag{2.30}$$

Variance of a random variable

The *variance* of a distribution is a measure of the *spread* of the distribution and it is often denoted by σ^2, in the scalar case, and by Σ in the multidimensional case. The variance of a continuous random variable X is defined as follows:

$$\begin{aligned}\text{var}[X] =&\mathbb{E}[(X-\mu)^2] = \int (x-\mu)^2 p(x)\mathrm{d}x = &(2.31)\\ &= \int x^2 p(x)\mathrm{d}x + \mu^2 \int p(x)\mathrm{d}x - 2\mu \int x \cdot p(x)\mathrm{d}x = \mathbb{E}[X^2] - \mu^2, &(2.32)\end{aligned}$$

from which it can be obtained the following relation:

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2. \tag{2.33}$$

From the variance of a rv, it is possible to define its *standard deviation* as:

$$\text{std}[X] = \sqrt{\text{var}[X]} = \sigma. \tag{2.34}$$

The variance of a rv has the following property:

$$\text{var}[aX + b] = a^2 \text{var}[X]. \tag{2.35}$$

Finally, given a set of n independent rvs, the variance of their sum is given by:

$$\text{var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{var}[X_i]. \tag{2.36}$$

Covariance of random variables

Given two discrete rvs X and Y possessing joint distribution $p(x,y)$, the corresponding *covariance* is defined as:

$$\begin{aligned}\mathrm{cov}[X,Y] =&\mathbb{E}_{X,Y}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = && (2.37)\\ =&\mathbb{E}_{X,Y}[XY]-\mathbb{E}[X]\mathbb{E}[Y] = && (2.38)\\ =&\sum_{x,y} x\cdot y\cdot p(x,y)-\sum_x x\cdot p(x)\sum_y y\cdot p(y), && (2.39)\end{aligned}$$

which represents the extent to which X and Y vary together. If X and Y are independent, then $\mathrm{cov}[X,Y]=0$, while if $Y=X$ it holds that $\mathrm{cov}[X,X]=\mathrm{var}[X]$. The continuous equivalent can be obtained easily by swapping sums with integrals.

Mode of a distribution

The *mode* of a distribution is the value associated with the highest probability mass or density:

$$x^* = \arg\max_x p(x). \tag{2.40}$$

A distribution is said to be *multimodal* when it has many local maxima, i.e. it has many different peaks.

Conditional Moments

When two or more dependent rvs are given it is possible to compute the moments of one given the knowledge about the others. The *law of iterated expectation*, also known as *law of total expectation* is defined as follows:

$$\mathbb{E}[X]=\mathbb{E}_Y[\mathbb{E}[X|Y]], \tag{2.41}$$

where the subscript Y in the first expectation refers to weighting the nested expectation by the distribution of Y. Such equation can be easily proven by exploiting the product rule (2.20) as follows:

$$\mathbb{E}_Y[\mathbb{E}[X|Y]] = \mathbb{E}_Y\left[\sum_x x \cdot p(x|y)\right] = \tag{2.42}$$

$$= \sum_y \left[\sum_x x \cdot p(x|y)\right] p(y) = \tag{2.43}$$

$$= \sum_{x,y} x \cdot p(x,y) = \mathbb{E}[X]. \tag{2.44}$$

The continuous equivalent is straightforward to obtain.

2.2 The Exponential Family

In the previous section probability fundamentals have been reported and the concept of probability distribution has been introduced from a general perspective; in this section the focus will be directed to a particular class of *parametric* distributions, namely the Exponential Family (EF) [17], which represents a broad set of models to describe the uncertainty of a process.

A parametric class of distribution is called an *exponential family* if it takes the following form:

$$p(x|\eta) = h(x)\, e^{\eta^T t(x) - a(\eta)}, \quad x \in \mathbb{R}^d, \tag{2.45}$$

where $t : \mathbb{R}^d \to \mathbb{R}^{n_\eta}$ is the vector of sufficient statistics, and $\eta \in \mathbb{R}^{n_\eta}$ is the vector of *natural* (or *canonical*) parameters, taking value in the natural parameter space Λ:

$$\Lambda = \left\{ \eta \in \mathbb{R}^{n_\eta} : 0 < \int h(x) \exp\left(\eta^T t(x)\right) dx < \infty \right\}. \tag{2.46}$$

Note: Λ can be proved to be an open and convex set. The support of $p(x|\eta)$ does not depend on η, and is encoded in the term $h(x)$. The function $a(\eta)$ is called the *log-normalizer*, and is such that

$$a(\eta) = \log(A(\eta)) = \log\left(\int h(x)\, e^{\eta^T t(x)} dx \right). \tag{2.47}$$

Let $b(\eta)$ denote the expected value of the sufficient statistics $t(x)$ for a given value of $\eta \in \Lambda$:

$$b(\eta) = E_\eta[t(x)]. \tag{2.48}$$

Well-known properties are

$$b(\eta) = \left(\frac{da(\eta)}{d\eta}\right)^T, \qquad \mathrm{cov}_\eta[t(x)] = \frac{d^2a(\eta)}{d\eta^2} > 0. \tag{2.49}$$

From the second of (2.49) it follows that the function $a(\eta)$ is strictly convex (note that $\mathrm{cov}_\eta[t(x)] \geq 0$ if the sufficient statistics $t(x)$ is not minimal).

The Bernoulli Distribution

The Bernoulli distribution is the discrete probability distribution of a random variable which takes the value 1 with probability α and the value 0 with probability $1-\alpha$. Less formally, it can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes–no question. Given the form (2.45), it has the following parameterization:

$$\begin{aligned}
&\text{Support: } x \in \{0,1\}\\
&\text{Parameter space: } \alpha \in [0,1]\\
&\eta = \eta_1,\\
&\eta_1 = \log\frac{\alpha}{1-\alpha},\\
&a(\eta) = \log(1+e^{\eta_1}),\\
&h(x) = 1,\\
&t(x) = x,
\end{aligned} \tag{2.50}$$

resulting in the following pdf:

$$\beta(x|\alpha) = \alpha^x(1-\alpha)^{1-x}. \tag{2.51}$$

The Poisson Distribution

In probability theory and statistics, the Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space, if these events occur with a known constant mean rate and independently of the time since the last event. Given the form (2.45), it has the following parameterization:

$$\begin{aligned}
&\text{Support: } x \in \mathbb{N}_0\\
&\text{Parameter space: } \lambda \in \mathbb{R}_+\\
&\eta = \eta_1,\\
&\eta_1 = \log \lambda,\\
&a(\eta) = e^{\eta_1},\\
&h(x) = \frac{1}{x!},\\
&t(x) = x,
\end{aligned} \tag{2.52}$$

resulting in the following pdf:

$$\pi(x|\lambda) = \frac{e^{-\lambda}}{x!}\lambda^x. \tag{2.53}$$

The Multivariate Gaussian Distribution

A core distribution which will be used in this work is the multivariate *Gaussian distribution* which, given the form (2.45), has the following parameters:

$$\begin{aligned}
&\text{Support: } x \in \mathbb{R}^d\\
&\text{Parameter space: } \mu \in \mathbb{R}^d,\ \Sigma \in S^d_{++}\\
&\eta = (\eta_1, \eta_2),\\
&\eta_1 = \Sigma^{-1}\mu,\\
&\eta_2 = -\frac{1}{2}\Sigma^{-1},\\
&a(\eta) = -\frac{1}{4}\eta_1^T\eta_2^{-1}\eta_1 - \frac{1}{2}\log|-2\eta_2|,\\
&h(x) = (2\pi)^{-d/2},\\
&t(x) = (x, xx^T).
\end{aligned} \tag{2.54}$$

Such parameterization provides the following pdf for the Gaussian case:

$$\nu(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, \tag{2.55}$$

where $x \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$ represents the mean of the Gaussian density, while $\Sigma \in S^d_{++} \subset \mathbb{R}^{d\times d}$ represents the corresponding covariance matrix. A more compact form for referring to Gaussian densities in this work will be $\nu(x)$.

The Gamma Distribution

Another distribution of interest for this work is the *Gamma* distribution, which has the following parameterization:

$$\begin{aligned}
&\text{Support: } \chi \in (0, +\infty) \\
&\text{Parameter space: } \kappa \in (0, +\infty),\ \omega \in (0, +\infty) \\
&\eta = (\eta_1, \eta_2), \\
&\eta_1 = \kappa - 1, \\
&\eta_2 = -\frac{1}{\omega}, \\
&a(\eta) = \log\Gamma(\eta_1 + 1) - (\eta_1 + 1)\log(-\eta_2), \\
&h(x) = 1, \\
&t(x) = (\log(x), x),
\end{aligned} \tag{2.56}$$

resulting in the following pdf:

$$\gamma(\chi|\kappa, \omega) = \frac{1}{\Gamma(\kappa)\omega^{\kappa}} \chi^{\kappa-1} e^{-\frac{\chi}{\omega}}, \tag{2.57}$$

where $\chi \in [0, \infty)$, $\kappa \in \mathbb{R}_+$ is the shape parameter, $\omega \in \mathbb{R}_+$ is the scale parameter and $\Gamma(\cdot)$ represents the *Euler Gamma function*, defined as:

$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} \mathrm{d}x. \tag{2.58}$$

The variable χ has been used in place of x to distinguish between the Gamma and the Gaussian density domains. In this work, a compact form for referring to gamma densities will be $\gamma(\chi)$

The Inverse Wishart Distribution

The last distribution here reported is the *inverse Wishart*, that is a probability distribution defined on real-valued positive-definite matrices, and which has the following parameterization:

$$
\begin{aligned}
&\text{Support: } \mathcal{Y} \in S^d_{++} \\
&\text{Parameter space: } v > 2d,\ V \in S^d_{++} \\
&\eta = (\eta_1, \eta_2), \\
&\eta_1 = -\frac{1}{2} v, \\
&\eta_2 = -\frac{1}{2} V, \\
&a(\eta) = \left(\eta_1 + \frac{d+1}{2}\right) \log|-\eta_2| + \log \Gamma_d\left(-\eta_1 - \frac{d+1}{2}\right), \\
&h(x) = 1, \\
&t(x) = (\log|x|, x^{-1}),
\end{aligned}
\tag{2.59}
$$

resulting in the following pdf:

$$
\varphi(\mathcal{Y}|V, v) = \frac{2^{-(\frac{v-d-1}{2})d} |V|^{\frac{v-d-1}{2}}}{\Gamma_d(\frac{\nu-d-1}{2}) |\mathcal{Y}|^{\frac{v}{2}}} e^{-\frac{1}{2}\operatorname{tr}(\mathcal{Y}^{-1}V)}, \tag{2.60}
$$

where $\mathcal{Y} \in S^d_{++}$, $v > 2d$ represents the scalar degrees of freedom (DoF) parameter, $V \in S^d_{++}$ is the matrix parameter, and $\Gamma_d(\cdot)$ is the *multivariate Euler Gamma function* defined as:

$$
\Gamma_d(z) = \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma\left(z + \frac{1-j}{2}\right). \tag{2.61}
$$

For other parameterization and details about distributions in the exponential family see [3, 18].

2.3 Maximum Likelihood Estimation

Let us consider a data set of N observations $\boldsymbol{x} = \{x_1, ..., x_N\}^T$, where $x_i \in \mathbb{R}^d$, $i = 1, ..., N$, drawn independently from a Gaussian distribution whose parameters are unknown; one can say that those samples are independent and identically distributed (i.i.d.). Since the data set is i.i.d., by exploiting (2.24), it is possible to write the probability of the data, given the parameters

μ and Σ of a Gaussian distribution, as:

$$p(\boldsymbol{x}|\theta) = p(x_1, ..., x_N|\mu, \Sigma) = \prod_{i=1}^{N} \nu(x_i|\mu, \Sigma), \tag{2.62}$$

where θ denotes the generic set of distribution parameters which, in the Gaussian case, are μ and Σ. When viewed as function of the parameters, the pdf is called *likelihood* function, since it assesses how likely are the drawn data points when specific parameters μ and Σ are given. If μ and Σ are assumed to be unknown, then maximizing (2.62) accordingly is a common criterion for determining the parameters of the pdf responsible for generating the observed data; such criterion is known as Maximum Likelihood Estimation (MLE) which, given a parametric distribution $p(x|\theta)$, and a set of i.i.d. observations $\boldsymbol{x} = \{x_1, ..., x_N\}$, is formally defined as:

$$\hat{\theta}_{ML} = \arg\max_{\theta} p(\boldsymbol{x}|\theta). \tag{2.63}$$

Note that this is a totally general criterion and it is not restricted to the Gaussian case. In the discussion above, it has been assumed that the observed data follows a Gaussian distribution, but in many problems such an assumption could be far from reality; when the distribution of the data is assumed, an operation of *model selection* is being performed, that is one is deciding which class the distribution underlying the observed data may belong to.

In practice, maximizing (2.62) can be difficult, and it is more common to consider the logarithm of the likelihood function. This is due to the fact that the log is a monotonically increasing function of its argument, hence the corresponding maximization is equivalent to maximize the function itself. Moreover, considering the logarithm not only simplifies the computations, but it even allows to deal with numerically small probabilities, so to avoid underflow events. Given that many optimization algorithms are designed to *minimize* instead of maximize, it will be considered the Negative Log-Likelihood (NLL), denoted with $\bar{\ell}(\boldsymbol{x}|\theta)$, so, by exploiting the logarithm properties, the problem becomes:

$$\hat{\theta}_{ML} = \arg\min_{\theta} \bar{\ell}(\boldsymbol{x}|\theta), \tag{2.64}$$

where:

$$\bar{\ell}(\boldsymbol{x}|\theta) = -\sum_{i=1}^{N} \log p(x_i|\theta). \tag{2.65}$$

MLE for Gaussian densities

In the Gaussian case, the corresponding formulation becomes:

$$\bar{\ell}(\boldsymbol{x}|\mu, \Sigma) = \frac{N}{2}\log|\Sigma| + \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu) + \frac{N}{2}d \cdot \log 2\pi. \quad (2.66)$$

In this case, the log-likelihood is strictly linked to the so called Square Mahalanobis' Distance (MD2); given a probability distribution $p(x) \in \mathbb{R}^d$, with mean μ and covariance matrix S, the MD2 of a point x from the distribution p is defined as:

$$D_{M2}(p\|x) = (x - \mu)^T S^{-1}(x - \mu). \quad (2.67)$$

Such distance can be seen as a multi-dimensional generalization of the idea of measuring how many standard deviations away x is from the mean of p. Given the NLL (2.66), one can write:

$$\bar{\ell}(\boldsymbol{x}|\mu, \Sigma) = \frac{1}{2}\sum_{i=1}^{N} D_{M2}(\nu\|x_i) + c(\Sigma), \quad (2.68)$$

where $c(\Sigma) = \frac{N}{2}\log|2\pi\Sigma|$, that is the NLL can be seen as the sum of the MD2s of all the samples from the Gaussian distribution $\nu(x|\mu, \Sigma)$. In order to minimize (2.66) w.r.t. μ and Σ, one can consider the following system:

$$\frac{\partial\bar{\ell}(\boldsymbol{x}|\mu, \Sigma)}{\partial\mu} = \sum_{i=1}^{N}\Sigma^{-1}(x_i - \mu) \overset{!}{=} 0, \quad (2.69)$$

$$\frac{\partial\bar{\ell}(\boldsymbol{x}|\mu, \Sigma)}{\partial\Sigma^{-1}} = -\frac{N}{2}\Sigma + \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T \overset{!}{=} 0, \quad (2.70)$$

which has solution:

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad (2.71)$$

$$\widehat{\Sigma}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{ML})(x_i - \hat{\mu}_{ML})^T. \quad (2.72)$$

$\hat{\mu}_{ML}$ is called *sample mean*, that is it corresponds to the arithmetic mean of the observed values, while $\widehat{\Sigma}_{ML}$ is the *sample covariance* calculated w.r.t. the sample mean.

Nonetheless, it is known that such criterion suffers of several issues, among which the underestimation of the true, yet unknown, distribution covariance due to a phenomena called *bias*, caused by the overfitting of the model to the observed data. A brief example of overfitting is the following: consider the scenario where one wants to predict the probability of heads when tossing a coin; after three experiments, three heads have been observed. The corresponding Maximum Likelihood (ML) estimate will be that the probability of a head is 1 while the chance of getting a tail is 0, which is rather unlikely. The equation for the unbiased covariance estimate is the following:

$$\widetilde{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu}_{ML})(x_i - \hat{\mu}_{ML})^T. \tag{2.73}$$

However, neither in this case the overfitting problem can be avoided. The effect of the bias becomes less impactful as the number of observed data points increases and, for $N \to \infty$, the ML covariance estimate is equal to the true distribution covariance. In Fig.2.1 is reported a plot where a set of i.i.d. samples generated according to $\nu(x|2,2)$ is fit by means of the ML principle (unbiased estimator):

MLE for gamma distributions

By proceeding as done for the Gaussian case, one can write the NLL for a set of N observations $\boldsymbol{\chi} = \{\chi_1, ..., \chi_n\}$ in the gamma case as follows:

$$\bar{\ell}(\boldsymbol{\chi}|\kappa,\omega) = (1-\kappa)\sum_{i=1}^{N} \log \chi_i + N \log \Gamma(\kappa) + N\kappa \log \omega + \frac{1}{\omega}\sum_{i=1}^{N} \chi_i \ . \tag{2.74}$$

By taking the derivative of the NLL w.r.t. ω, and by equating it to zero, one obtains the following equation:

$$\omega = \frac{1}{\kappa} \cdot \frac{1}{N} \sum_{i=1}^{N} \chi_i \ . \tag{2.75}$$

By doing the same for κ, and by substituting the result (2.75), one obtains:

$$\psi_0(\kappa) - \log \kappa = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \log \chi_i - \log \frac{1}{N}\sum_{i=1}^{N} \chi_i}_{\triangleq g} . \tag{2.76}$$

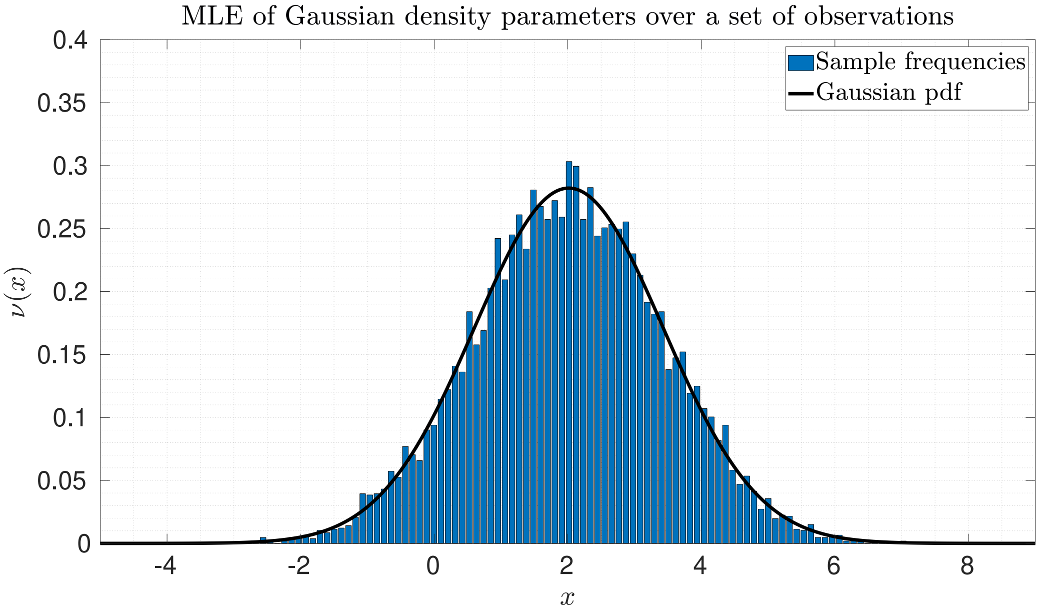

Figure 2.1: Set of $N = 10000$ observations frequencies regrouped in bins (blue), estimated Gaussian pdf (solid black). The estimated parameters are $\hat{\mu}_{ML} = 2.0000$, $\tilde{\Sigma}_{ML} = 2.0264$.

where $\psi_0(\kappa)$ is the *digamma* function (a.k.a. *polygamma* of order zero) which corresponds to the first order derivative of the logarithm of the Euler gamma function $\Gamma(\cdot)$. The equation (2.76) does not admit an analytic solution; in this regard, one can resort to the Newton-Rhapson (NR) algorithm as follows:

$$f(\kappa) \triangleq \psi_0(\kappa) - \log \kappa - g, \tag{2.77}$$

$$f'(\kappa) = \psi_1(\kappa) - \frac{1}{\kappa}, \tag{2.78}$$

where g has been defined in (2.76), $\psi_1(\kappa)$ is the *trigamma* function (a.k.a. polygamma function of order one) and $f'(\kappa)$ is the derivative of $f(\kappa)$ w.r.t. κ.

By choosing now an initial (positive) value for κ, say $\kappa^- > 0$, one can write the following recursion:

$$\kappa^+ = \kappa^- - \frac{f(\kappa^-)}{f'(\kappa^-)}. \tag{2.79}$$

The iterations can stop after either a maximum allowed number or when the variation in the update falls below a given threshold. The final value will be $\hat{\kappa}_{ML}$. Once $\hat{\kappa}_{ML}$ is available, one can substitute it back to (2.75) to obtain $\hat{\omega}_{ML}$; an accurate starting point for the NR algorithm is $\kappa_0 = -\frac{1+\sqrt{1-4g/3}}{4g}$, which can be obtained by exploiting the *generalized Puisex series* to approximate the left-hand side of (2.76).

In Fig.2.2 is reported a plot where a set of i.i.d. samples generated according to $\gamma(\chi|7,1)$ is fit by means of the ML principle:

MLE for inverse-Wishart distributions

The inverse-Wishart inherits the same argumentation proposed for the gamma case. The corresponding NLL for a set of i.i.d samples $\boldsymbol{\mathcal{Y}} = \{\mathcal{Y}_1, ..., \mathcal{Y}_N\}$ (SPD matrices) is as follows:

$$\begin{aligned}\bar{\ell}(\boldsymbol{\mathcal{Y}}|V, v) =& N\left(\frac{v-d-1}{2}\right) d \log 2 + N\left(\frac{v-d-1}{2}\right) \log|V| + \\ & -N \log \Gamma_d\left(\frac{v-d-1}{2}\right) - \frac{v}{2}\sum_{i=1}^{N} \log|\mathcal{Y}_i| - \frac{1}{2}\mathrm{tr}(\sum_{i=1}^{N} \mathcal{Y}_i^{-1} \cdot V).\end{aligned} \tag{2.80}$$

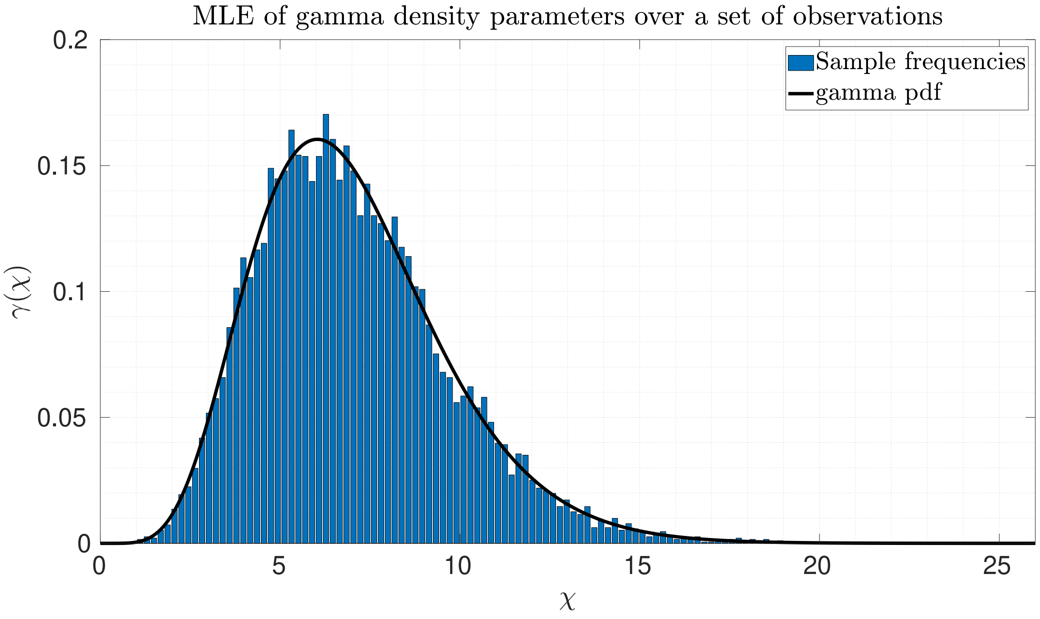

Figure 2.2: Set of $N = 10000$ observations frequencies regrouped in bins (blue), estimated gamma pdf (solid black). The estimated parameters are $\hat{\kappa}_{ML} = 7.0090$, $\hat{\omega}_{ML} = 0.9947$.

By deriving w.r.t. V and equating to zero, one gets the following equation:

$$V = (v-d-1)\underbrace{\left[\frac{1}{N}\sum_{i=1}^{N}\mathcal{Y}_i^{-1}\right]^{-1}}_{\triangleq H_1}. \tag{2.81}$$

The term H_1 is equivalent to the *geometric* mean of the samples. The NLL derivative w.r.t. v yields instead:

$$\sum_{j=1}^{d}\psi_0\!\left(\frac{v-d-j}{2}\right) - d\log\left(\frac{v-d-1}{2}\right) = \log|G_1| - \underbrace{\frac{1}{N}\sum_{i=1}^{N}\log|\mathcal{Y}_i|}_{\triangleq h_2}, \tag{2.82}$$

where the first term in the left-hand side corresponds to the derivative of the multivariate Euler gamma function. Resorting again to the NR algorithm, one considers:

$$f(v) \triangleq \sum_{j=1}^{d}\psi_0\!\left(\frac{v-d-j}{2}\right) - d\log\left(\frac{v-d-1}{2}\right) - \log|H_1| + h_2, \tag{2.83}$$

$$f'(v) = \frac{1}{2}\sum_{j=1}^{d}\psi_1\!\left(\frac{v-d-j}{2}\right) - \frac{d}{v-d-1}, \tag{2.84}$$

and $\hat{v}_{ML}$ can again be estimated by means of the recursion:

$$v^{+} = v^{-} - \frac{f(v^{-})}{f'(v^{-})}. \tag{2.85}$$

Once $\hat{v}_{ML}$ is available, one can substitute it back in (2.81) to obtain $\widehat{V}_{ML}$. As discussed for the gamma case, even for the inverse-Wishart parameter v it is necessary to pick a suitable initial value to start the NR recursion; given the parameterization considered in this work, there is a lower bound in such choice, that is one has to pick $v_0 > 2d$. As discussed in [19] for the barycenter case, a suitable value can be $v_0 = 2d + \epsilon$, where ϵ is a small enough value, e.g., 10^{-5}, which in general avoids the NR algorithm to explore solutions $v^{+} < 2d$. Both for the gamma and inverse-Wishart case, the ML estimate exists unique, hence the NR initialization should just guarantee the mentioned requirements. Regarding the sampling of an inverse-Wishart distribution, one can use the MATLAB command *iwishrnd*. Nonetheless, in order to generate samples $\boldsymbol{\mathcal{Y}}$ from which it is possible to estimate the exact chosen parameters V and v by using the equations here provided, one should generate samples by considering $iwishrnd(V, v-d-1)$. In Fig .2.3 is reported a plot where a set of i.i.d. samples generated according to $\varphi(\mathcal{Y}|5, 20)$ is fit by means of the ML principle.

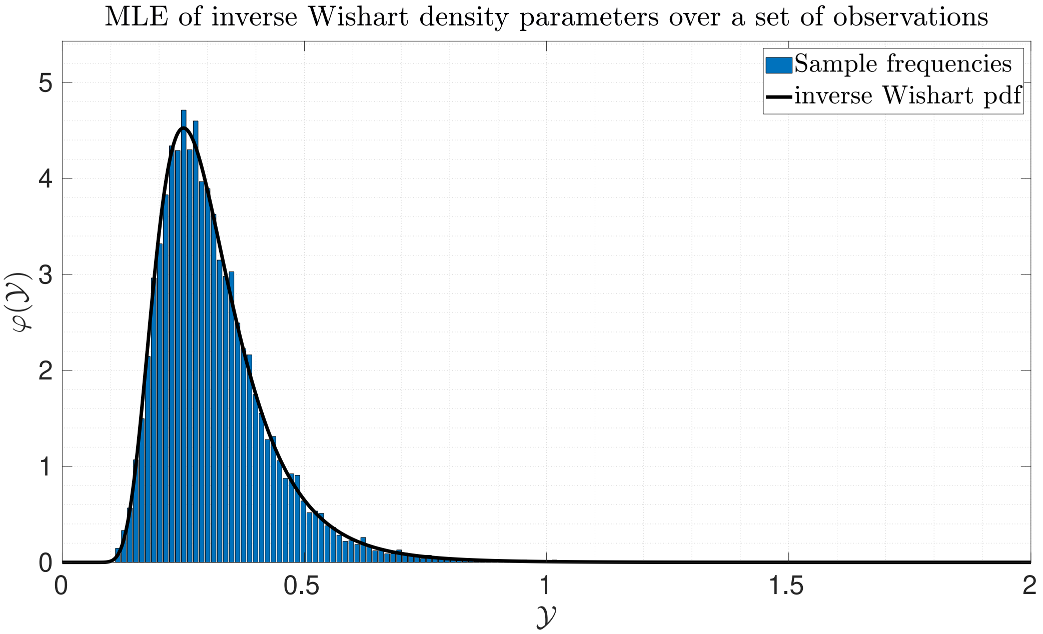

Figure 2.3: Set of $N = 10000$ observations frequencies regrouped in bins (blue), estimated inverse Wishart pdf (solid black). The estimated parameters are $\hat{V}_{ML} = 5.0275$, $\hat{v}_{ML} = 20.2014$.

Additional notes

In the section 2.5 the Bayesian approach will be discussed as opposite of the frequentist statistics, where no prior information is considered in the estimation process. Maximum likelihood is a *frequentist approach*, since the corresponding estimates do not rely on prior information but only on the observed data, hence small sample sets can lead to a high bias, as mentioned. Nonetheless, it can at the same time be motivated in a Bayesian perspective by considering a *non-informative* or *uniform* prior distribution. More on this will be discussed in the following sections.

2.4 Maximum A Posteriori Estimation

As mentioned, the MLE suffers from the phenomenon of overfitting. The core problem is that the chosen model has enough parameters to fit perfectly the observed data, hence it can perfectly match the empirical distribution (built solely on the data). Nonetheless, the empirical distribution is not the same as the true one, so fitting tightly the observed set of samples will not leave over any probability for novel data. In other terms, the model may not be able to *generalize*. If an *a priori* knowledge $p(\theta)$ is available about the parameter(s) of interest θ, by exploiting the Bayes' rule (2.21), it is possible to treat it as a random variable, hence to compute its posterior distribution as:

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int p(\boldsymbol{x}|\theta)p(\theta)\mathrm{d}\theta}. \tag{2.86}$$

First of all, the denominator in (2.86) does not depend on θ, but it serves as a normalization constant to guarantee that the resulting posterior will be a pdf, hence integrating to one. If the most probable value for the parameter(s) is sought, then one can solve the following problem:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\boldsymbol{x}|\theta)p(\theta). \tag{2.87}$$

Such problem falls under the name of Maximum a Posteriori (MAP) estimation. By doing the same considerations as for the MLE case, one can consider the negative logarithm of the problem in order to obtain:

$$\hat{\theta}_{MAP} = \arg\min_{\theta} \left[\bar{\ell}(\boldsymbol{x}|\theta) - \log p(\theta)\right]. \tag{2.88}$$

The second term in (2.88) can be seen as a *regularization* term which controls the complexity of the resulting model. In general, regularization is a practice employed in many estimation problems in order to alleviate the effect of overfitting. By recalling now the problem of estimating the probability of landing, for instance, a tail while tossing a coin, it is clear that the formulation (2.88) is more robust; if the a priori distribution assumes that the coin is fair, i.e. the probabilities of landing either a tail or a head are equal to 0.5, then even by observing three tails straight would not provide the unlikely estimate assigning all the probability to the event of observing a tail, as happened for the MLE case. At the opposite, one is more likely to deviate from the "fair coin" case, but still leaving a chance to obtain a head in the fourth tossing. Although the MAP can be seen as a limit case of the Bayesian estimation, it is not very representative of such methods in general, since the corresponding estimates are point estimates, whereas Bayesian estimates provide the whole distribution of the parameter(s) of interest. Moreover, in the Bayesian case it is more common to report the posterior mean or median as point estimate, together with confidence intervals, instead of the most probable value (in some cases it coincides with the distribution mean, e.g., the Gaussian case).

2.5 Bayesian Estimation

Until now, several ways to estimate parameters have been discussed. Nonetheless, all of them provide point estimates and ignore the corresponding uncertainty, which in general represents a high source of information in many applications. In statistics, using a probability distribution to model the uncertainty about a parameter is known as *inference.* More in detail, in the Bayesian framework, the distribution modelling the uncertainty corresponding to a quantity of interest is the *posterior* distribution, which is the combination between the prior distribution, that is the available knowledge before observing new data, and the likelihood function, that is the data which is expected to be observed for each possible value of the parameter of interest. By recalling the Bayes formula (2.21), given a set of observations $\boldsymbol{x}$, the posterior distribution of the parameter(s) θ can be computed as:

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int p(\boldsymbol{x}|\theta)p(\theta)\mathrm{d}\theta}, \tag{2.89}$$

where the term $p(\boldsymbol{x})$ is the *marginal likelihood* (or *evidence*) since it is obtained by marginalizing over the unknown parameter θ. In general, the denominator of (2.89) is constant and independent of θ, and it serves as a normalization constant to guarantee the constraint of pdfs that $\int p(\theta|\boldsymbol{x})\mathrm{d}\theta = 1$.

If compared to the ML or MAP estimates, the Bayesian inference yields, as mentioned, a distribution rather than a point estimate, hence the uncertainty about the true value of the parameter plays a role in the estimation process. If one has to extract a point estimate of the parameter θ, a common approach is to evaluate either the mean of the posterior distribution or the corresponding median.

Conjugate priors

A prior distribution $p(\theta)$ belonging to a family $\mathcal{Q}$ (e.g., Gaussian, Gamma, inverse-Wishart etc.) is said to be a *conjugate prior* for a likelihood function $p(\boldsymbol{x}|\theta)$, if the posterior distribution is in the same class of the prior, that is $p(\theta|\boldsymbol{x}) \in \mathcal{Q}$. If the class $\mathcal{Q}$ falls inside the exponential family (see 2.2), then the posterior distribution can be computed in closed form. The Gaussian distribution (2.55) is the conjugate prior of itself, in the sense that given a Gaussian prior, and a Gaussian likelihood function, then the posterior will be itself Gaussian. The inverse-Wishart (2.60) represents a conjugate prior for the covariance matrix of a multivariate Gaussian distribution, while the Gamma distribution (2.57) represents a conjugate prior for the rate of a Poisson distribution [20]. Further details on conjugacy are outside the goals of this work, but the three mentioned distributions will be discussed more in detail when the topic of GGIW (gamma Gaussian inverse Wishart) reduction will be addressed in Chapter 5.

2.6 Fundamentals of Information Theory

Let us consider a discrete random variable X distributed according the probability distribution $p(x)$. How much *information* is received when a specific realization of this variable is observed? The amount of information can be viewed as the *degree of surprise* when an observation of the values X can take is done. If a highly improbable event has just occurred, it would be reasonable to assume that the received information is higher if compared to the case where an almost certain event has happened. The information content will therefore depend on the probability distribution $p(x)$; in order to evaluate the information content, it can be convenient to consider a monotonic function $h(\cdot)$ of $p(x)$. Which form should $h(\cdot)$ take? If one considers two unrelated events x and y, then it would be reasonable to say that the information gain from observing both should be the sum of the information gained from observing those separately, that is it should be $h(x, y) = h(x) + h(y)$. Two unrelated events are statistically independent, that is $p(x, y) = p(x)p(y)$.

One suitable candidate for $h(x)$ is the base-2 *logarithm* function, that is:

$$h(x) = -\log_2 p(x), \tag{2.90}$$

where the negative sign guarantees that information is either positive or zero. The choice of basis for the logarithm is arbitrary, but for now the base 2 is considered, which corresponds to the units of *binary digits (bits).* Suppose now that a sender wishes to transmit the value of a random variable to a receiver. Then, the average amount of transmitted information in the process is obtained by taking the average of (2.90) w.r.t. $p(x)$, that is:

$$H[x] = -\sum_x p(x) \log_2 p(x). \tag{2.91}$$

where $H[x]$ is a compact form to denote either $H[p(x)]$ or $H[X]$. The above quantity is called *entropy* of the random variable X. Although a heuristic motivation for the quantities above has been provided, it results that it finds important uses in many of the modern applications. For instance, in 1948 Shannon [21] stated in the *noiseless coding theorem* that the entropy represents a lower bound on the number of bits needed to transmit the state of a random variable.

If the natural logarithm, which will be denoted simply as $\log(\cdot)$, is instead considered in (2.90), the unit of information is *nats* instead of *bits.* Note that distributions $p(x)$ that are sharply peaked around few values will have a relatively low entropy, whereas those that are spread more evenly across many values will have higher entropy. Until now, the discrete case for probability distributions has been considered, for which the entropy is guaranteed to be non-negative. Many applications though, have to deal with continuous random variables; in this regard, it is possible to extend the concept of entropy to continuous distributions [3], which provides the so called *differential entropy*, defined as follows:

$$H[x] = -\int p(x) \log p(x) \mathrm{d}x. \tag{2.92}$$

Interchangeably, one can write $H[p]$ to refer to the entropy of a random variable distributed according to $p(x)$. The main difference of the *differential* entropy (2.92) defined for continuous rv from the entropy defined in (2.91) for discrete rv is that the differential entropy can also take negative values. It can be proved that, given a continuous rv x with given mean μ and covariance Σ, the distribution that has the maximum differential entropy is the Gaussian distribution. The entropy of a Gaussian of mean μ and covariance Σ has the following closed form:

$$H[\nu] = \frac{1}{2}(\log|\Sigma| + d\log 2\pi + d), \tag{2.93}$$

which, therefore, is the largest entropy that can have a rv with mean μ and covariance Σ. For the gamma case, the entropy is:

$$H[\gamma] = \kappa + \log\Gamma(\kappa) + \log\omega + (1-\kappa)\psi_0(\kappa), \tag{2.94}$$

where $\psi_0(\kappa)$ is the *digamma* function (a.k.a. *polygamma* function of order zero).
For the inverse-Wishart case, the entropy takes the following form:

$$H[\varphi] = \log\Gamma_d\bigg(\frac{v-d-1}{2}\bigg)+\frac{v-d-1}{2}d+\frac{d+1}{2}\log\left|\frac{1}{2}V\right|-\frac{v}{2}\sum_{j=1}^{d}\psi_0\bigg(\frac{v-d-j}{2}\bigg). \tag{2.95}$$

Suppose now that a joint distribution $p(x, y)$ is available from which pairs of values of x and y are drawn. If a value of x is already known, then the additional information needed to specify the corresponding value of y is given by $-\log p(y|x)$. Hence, the average additional information needed to specify y can be written as:

$$H[y|x] = -\iint p(x, y)\log p(y|x)\mathrm{d}x\mathrm{d}y. \tag{2.96}$$

Such a quantity is called the *conditional entropy* of y given x. By using the product rule, that is $p(y|x) = p(x, y)/p(x)$, one obtains:

$$H[x, y] = H[y|x] + H[x], \tag{2.97}$$

where $H[x, y]$ is the differential entropy of $p(x, y)$ and $H[x]$ is the differential entropy of the marginal distribution $p(x)$. Thus, the information needed to describe x and y is given by the sum of the information needed to describe x alone plus the additional information required to specify y given x.

Consider now an unknown distribution $p(x)$ from which is possible to draw samples, and suppose that a model $q(x)$ is used to construct an efficient coding scheme for the purpose of transmitting values of x to a receiver. It is then possible to define the following quantity:

$$H_\times[p,q] = -\int p(x)\log q(x)\mathrm{d}x. \tag{2.98}$$

The equation above defines the so called differential *cross-entropy*, which can be showed to be the average amount of information which is required to encode the observation relative to $p(x)$ using a coding scheme based on $q(x)$ [16]. More in general, the cross-entropy is used as a *loss function* to determine how well a chosen model $q(x)$ is fitting data samples drawn from the unknown $p(x)$. In the Gaussian case, the cross-entropy between $\nu_i = \nu(x|\mu_i,\Sigma_i)$ and $\nu_j = \nu(x|\mu_j,\Sigma_j)$ takes the following form:

$$H_\times[\nu_i,\nu_j] = \frac{1}{2}\big[d\log 2\pi + \mathrm{tr}(\Sigma_j^{-1}\Sigma_i) + (\mu_i-\mu_j)^T\Sigma_j^{-1}(\mu_i-\mu_j) + \log|\Sigma_j|\big]. \tag{2.99}$$

In the gamma case, the cross-entropy between $\gamma_i(\chi) = \gamma(\chi|\kappa_i,\omega_i)$ and $\gamma_j(\chi) = \gamma(\chi|\kappa_j,\omega_j)$ takes the following form:

$$H_\times[\gamma_i,\gamma_j] = (1-\kappa_j)[\log\omega_i + \psi_0(\kappa_i)] + \log\Gamma(\kappa_j) + \kappa_j\log\omega_j + \kappa_i\frac{\omega_i}{\omega_j}. \tag{2.100}$$

In the inverse-Wishart case, the cross-entropy between $\varphi_i(\mathcal{Y}) = \varphi(\mathcal{Y}|V_i,v_i)$ and $\varphi_j(\mathcal{Y}) = \varphi(\mathcal{Y}|V_j,v_j)$ takes the following form:

$$\begin{aligned} H_\times[\varphi_i,\varphi_j] =& \frac{d+1}{2}\log\left|\frac{1}{2}V\right| + \log\Gamma_d\left(\frac{v_j-d-1}{2}\right) + \frac{v_i-d-1}{2}\mathrm{tr}\big(V_i^{-1}\cdot V_j\big) + \\ & -\frac{v_j-d-1}{2}\log\big|V_i^{-1}\cdot V_j\big| - \frac{v_j}{2}\sum_{m=1}^{d}\psi_0\left(\frac{v_i-d-m}{2}\right). \end{aligned} \tag{2.101}$$

2.7 Model Selection

As mentioned in the previous sections, when trying to model the uncertainty of a process of interest one can *select* a parametric statistical model and try to estimate the corresponding parameters which better fit the observed data. The choice of such a representation, though, can be really impactful if one wants to employ it in tasks like making predictions or computing specific statistics about the observed process; performing a *model selection* is in general a really difficult task, since the true behavior of a system could be far from the guessed model, or it could be complex enough that no available standard model can describe it accurately. Given some data, in order to maximize the likelihood on a model of such data, one can increase the number of parameters used, but by doing so the phenomenon of overfitting may arise. In this regard, it would be useful to have some criteria which can identify a suitable trade-off between the likelihood of a model over the data and the corresponding amount of parameters. In this section, some fundamentals about model selection criteria will be reported.

Given a set of observations $\boldsymbol{x} = \{x_1, ..., x_N\}$, there exists several criteria which allow to evaluate how well a given model represents such data. To make the following argumentation more clear, let us consider a d-dimensional Gaussian density; such a model will have in total $\tilde{m} = d + \frac{d(d+1)}{2}$ independent parameters, with $\tilde{m}$ order of the model, since the mean vector will possess d components, while the d-dimensional covariance matrix can be described by considering only the upper triangular part, since symmetric, hence having in total $\frac{d(d+1)}{2}$ independent parameters. Let us now consider a generic parametric model $p(x|\theta)$ having $\tilde{m}$ independent parameters ($\theta = \theta(\tilde{m})$); the Akaike Information Criterion (AIC) is defined as follows:

$$AIC = -2 \log p(\boldsymbol{x}|\theta) + 2\tilde{m} = 2\bar{\ell}(\boldsymbol{x}|\theta) + 2\tilde{m}. \tag{2.102}$$

Such a criterion quantifies how the $\tilde{m}$-parameter model fits over the observed N samples; the smallest the value, the better the trade-off between the number of parameters and the corresponding likelihood over the data. Hence, if one wants to evaluate how well a model fits the observed data, should always select, according to this criterion, the model yielding the smallest AIC value. Nonetheless, such a criterion is known to penalize insufficiently the model complexity, hence yielding, in general, more complex models if compared to the Bayesian Information Criterion (BIC), defined as follows:

$$BIC = -2 \log p(\boldsymbol{x}|\theta) + \tilde{m} \log N = 2\bar{\ell}(\boldsymbol{x}|\theta) + \tilde{m} \log N. \tag{2.103}$$

As for the AIC, the model order associated with the smallest value of the BIC will represent the best trade-off between the number of parameters and the fitting of the model over the data. For a more detailed discussion about the criteria discussed above, see [22].

Although those criteria appear to be simple, their utilization might be hindered by several factors; for instance, assume that given a set of observations, one estimates the parameter of a model with order $\tilde{m}$ over such data. For a given order, the way the parameters are estimated can in general influence the resulting value of the information criterion: to the same order can be associated different models which may differ significantly. In addition, if more complex models are used, e.g., mixture of densities (discussed in the next section), even by considering the same parameter estimation method, for the same model order, there can exist different realizations. Thus said, when modelling phenomena which may exhibit complex behaviors, the task of model selection becomes even more difficult. Another hindrance of the information criteria as above, is the fact that one should evaluate several model orders over the same data to obtain a potentially suitable representation; this might be computationally prohibitive for many applications where a large amount of high-dimensional data has to be processed.

2.8 Mixtures of densities

The material following in this section is totally general, but often Gaussianity will be used for the sake of discussion; nonetheless, all the presented methodology can be applied for, at least, the whole exponential family. As mentioned, many real world problems can exhibit complex behaviors (e.g., multimodality) which simple models can not approximate accurately. In those cases, a possible approach can be to consider combinations of simpler models, in order to provide more accurate representations, as follows:

$$p(x|\Theta) = \boldsymbol{w}^T \boldsymbol{q}(x|\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i q(x|\theta_i) = \sum_{i=1}^{n} w_i q_i, \tag{2.104}$$

where $\boldsymbol{q}$[1] is a vector of generic parametric pdfs $q(x|\theta_i) \in \mathcal{Q}$, e.g., $\boldsymbol{q} = [q_1, ..., q_n]^T$, belonging to the same family $\mathcal{Q}$ (e.g., Gaussians, gammas, etc.), $\boldsymbol{w} = [w_1, ..., w_n]^T$ is a vector of *weights* with $w_i \in [0, 1], \boldsymbol{w}^T \mathbb{1}_n = 1$. $p(x|\Theta) \in$

[1] With a slight abuse of notation, throughout this work $\boldsymbol{q}$ will be sometimes interpreted as a vector, sometimes as a set, whose components are the n distributions q_i, $i \in [1:n]$, in a family of distributions $\mathcal{Q}$. As a set, $\boldsymbol{q} = \{q_i\}_{i=1}^n \subset \mathcal{Q}$, as a vector $\boldsymbol{q} \in (\mathcal{Q})^n$. The interpretation of $\boldsymbol{q}$ as a vector allows us to write a mixture in the compact form $\boldsymbol{w}^T\boldsymbol{q}$ instead of $\sum_{1=1}^n w_i q_i$. Also the vector $\boldsymbol{w}$ of weights will be often interpreted as a set $\{w_i\}_{i=1}^n \subset [0, 1]$.

$\mathcal{Q}_{\text{mix}}$ is a Mixture of Densities (MoDs) of size n obtained as the *convex sum* of the q_i components, where $\mathcal{Q}_{\text{mix}}$ denotes the space of all mixture of densities with components in $\mathcal{Q}$; sometimes it will be used the symbol $\mathcal{Q}_{\text{mix}}^{(n)}$, which denotes the space of all the mixtures belonging to the family $\mathcal{Q}$ of exactly size n. Convexity in the sum is required in order to preserve the pdf constraint $\int p(x|\Theta)\mathrm{d}x = 1$, that is a mixture of densities is a density itself. $\Theta = \{\boldsymbol{w}, \boldsymbol{\theta}\} \in \mathcal{H}_n = \Delta^{n-1} \times \mathcal{H}_n^\theta$ is the collection of all the parameters of the mixture which depends on the family considered, with $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^n$, where $\mathcal{H}_n^\theta$ is the space over which all the parameters, but the weights, of the mixture components are defined, and $\Delta^{n-1} = \{\boldsymbol{w} \in \mathbb{R}_+^n : \ \boldsymbol{w}^T \mathbb{1}_n = 1\} \subset \mathbb{R}_+^n$ is the standard simplex of dimension $n-1$; for instance, in the Gaussian case $\Theta = \{\boldsymbol{w}, \boldsymbol{\theta}\} = \{\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, where $\boldsymbol{w} \in \mathbb{R}^n$ is the weight vector, $\boldsymbol{\mu} \in (\mathbb{R}^d)^n$ is the collection of all the means, while $\boldsymbol{\Sigma} \in (S_{++}^d)^n$ is the collection of all the covariance matrices. If the convexity of the sum is removed, then the *un-normalized* weighted sum of densities is called *intensity*; to distinguish from mixtures, the space of the intensities will be denoted as $\mathcal{Q}_{\text{int}}$. For the remainder of this work, though, unless specifically said, all the argumentation will be done in terms of mixtures rather than intensities. For a matter of simplicity, when dealing with mixtures, the model order will often be considered to be $m = n$; nonetheless, in the perspective presented in sec. 2.7, the real model order would be $m = n \cdot \tilde{m}$, where $\tilde{m}$ is the amount of free parameters per component.

Definition 2.8.1. *(Sub-mixtures) Consider a mixture $p(x|\Theta)$ of size n, with parameter set $\Theta = \{\boldsymbol{w}, \boldsymbol{\theta}\}$, and consider a set of indexes, $\mathcal{I} \subset [1\colon n]$ of size $\bar{n} < n$, used to select a subset of Θ, denoted $\Theta_{\mathcal{I}} = \{\boldsymbol{w}_{\mathcal{I}}, \boldsymbol{\theta}_{\mathcal{I}}\} \subset \Theta$. Then, $p(x|\Theta_{\mathcal{I}})$ is a <u>un-normalized</u> sub-mixture of $p(x|\Theta)$ ($\boldsymbol{w}_{\mathcal{I}}^T \mathbb{1}_{\bar{n}} < 1$ and $\boldsymbol{w}_{\mathcal{I}} \notin \Delta^{\bar{n}-1}$). The set of parameters $\bar{\Theta}_{\mathcal{I}} = \{\bar{\boldsymbol{w}}_{\mathcal{I}}, \boldsymbol{\theta}_{\mathcal{I}}\}$, where $\bar{\boldsymbol{w}}_{\mathcal{I}} = \boldsymbol{w}_{\mathcal{I}}/(\boldsymbol{w}_{\mathcal{I}}^T \mathbb{1}_{\bar{n}})$, defines a <u>normalized</u> sub-mixture of $p(x|\Theta)$ (i.e., $\bar{\boldsymbol{w}}_{\mathcal{I}} \in \Delta^{\bar{n}-1}$).*

Definition 2.8.2 (Degenerate mixture models)**.** *A mixture model $p(x|\Theta)$ of size n is said to be <u>degenerate</u> if $\bar{n} < n$ of its components are equal; in that case, the mixture components associated with the corresponding set of indices $\mathcal{I} \subset [1:n]$, of size $\bar{n} < n$, can be replaced by a single density having weight $w_{\mathcal{I}} = \sum_{\mathcal{I}} w_i$ and parameters equal to one of the components in the considered partition.*

Definition 2.8.3 (Singular mixture models)**.** *Given a mixture model $p(x|\Theta)$ of size n, it is said to be <u>singular</u> if all of its components are equal; in that* *case, the mixture model can be replaced by a single density having weight* $w = 1$ *and parameterized as one of the mixture components.*

Why mixtures of densities?

Problems like localizing a robot in a room [2], tracking targets under the presence of clutter [12], filtering of nonlinear/switching stochastic systems [4], unsupervised learning of statistical models [9] and so on, have to deal often with multimodal uncertainties. In those cases, estimating the true distribution analytically may result to be very difficult, if not impossible; moreover, if the computational capabilities are limited, the corresponding approximation should be efficient in terms of employed resources. In this regard, mixtures of parametric densities may be able to both preserve the representation accuracy and save computational resources via an efficient mathematical description. As discussed in [12, 23], Gaussian mixtures can approximate with arbitrary accuracy any kind of density; in those works, the problem of bayesian filtering is addressed for nonlinear stochastic systems, but such considerations are way more general. In addition, in many recent approaches to target tracking [20, 24–27], intensities are used to estimate the kinematic state, the extent or other features of point/extended objects in the presence of clutter.

Mixture densities are, then, a widely used tool given that they represent an efficient, versatile, yet powerful, tool to describe complex behaviors or features. Nonetheless, especially in Bayesian estimation contexts, when the distribution of interest is approximated by a mixture, the corresponding number of components can become very large, leading to a computationally intractable representation after few iterations.

Additional notes on mixture representations

The following discussion holds for any of the distributions in the exponential family; nonetheless, the Gaussian case will be addressed for the sake of argumentation.

Let us introduce a n-dimensional binary random variable $Y = [y_1, ..., y_n]^T$ having a 1-of-n representation, that is if $y_i = 1$ all the other elements are zero, hence there are n possible states for Y; moreover, it holds that $y_i \in \{0, 1\}$ and $\sum_{i=1}^{n} y_i = 1$. In addition, let us write the distribution $p(Y)$ in terms of mixing coefficients w_i, such that:

$$p(y_i = 1) = w_i, \tag{2.105}$$

where $\sum_i w_i = 1$, $w_i \in [0, 1]$. Given that Y has a 1-of-n representation, one can write its overall distribution (pdf) as:

$$p(Y) = \prod_{i=1}^{n} w_i^{y_i}. \tag{2.106}$$

Let us now assume that the data of interest has been generated according to a Gaussian random variable X. The corresponding conditional distribution, assuming a specific realization of the *latent*[2] variable Y, can then be written as:

$$p(X = x|y_i = 1) = \nu(x|\mu_i, \Sigma_i), \tag{2.107}$$

with the overall conditional distribution defined as:

$$p(x|Y) = \prod_{i=1}^{n} \nu(x|\mu_i, \Sigma_i)^{y_i}. \tag{2.108}$$

By exploiting the product rule (2.20), it is then possible to write the joint distribution of the data and the latent variable as:

$$p(x, Y) = p(x|Y)p(Y). \tag{2.109}$$

Marginalizing now over all the possible realizations of Y, one can obtain the distribution of the observed data as:

$$p(x) = \sum_{y_i} p(x|Y = y_i)p(Y = y_i) = \sum_{y_i} p(x, Y = y_i) = \sum_{i=1}^{n} w_i \nu(x|\mu_i, \Sigma_i). \tag{2.110}$$

Thus, the marginal distribution $p(x)$ is a Gaussian mixture. Moreover, for each observation x_j, $j = 1, ..., N$, one can suppose that there is a corresponding latent variable y_i which is *responsible* for its generation. In this regard,

[2] Latent, or hidden, variables are variables that are not directly observed, but instead can be inferred through a mathematical model describing the potential relations with other observable variables.

let us consider the following quantity:

$$\begin{aligned}\rho(y_i) \triangleq p(y_i = 1|X = x) \overset{Bayes}{=} & \frac{p(X = x|y_i = 1)p(y_i = 1)}{\sum_{k=1}^{n} p(X = x|y_k = 1)p(y_k = 1)} \\ = & \frac{w_i \nu(x|\nu_i, \Sigma_i)}{\sum_{k=1}^{n} w_k \nu(x|\mu_k, \Sigma_k)}.\end{aligned} \tag{2.111}$$

The quantity $\rho(y_i)$ is called *responsibility*, and it represents the probability that one of the n possible Gaussian distributions is responsible for the observed data sample; the weights w_i should be seen as the probability of $y_i = 1$.

Ancestral sampling: generating data from mixture models

In order to generate a set of observations $\boldsymbol{x} = \{x_1, ..., x_N\}$ from a mixture model, one can proceed as follows:

1. At first, one the model is given, hence the distribution of the latent variable Y is known, that is one can obtain a realization of one of the n possible corresponding configurations. This step will select which component in the mixture is "active", by excluding all the others.

2. Given a realization $Y = y_i$, it is then possible to sample from the drawn density corresponding to the distribution $p(x|Y = y_i)$.

3. Repeat (1) and (2) until the desired amount of samples has been drawn.

How to draw samples from $p(Y)$? As mentioned, the weights w_i represent the probability of realizing one out of the n schemes of the latent variable; in this regard, one can at first compute the discrete cumulative distribution function $P(Y)$ of such weights, that is, starting from the weight associated to y_1, the cdf is obtained by cumulatively summing up all the weights up to the one of y_n. The weights add up to one, hence the cdf will be a curve starting from zero and ending in one. By drawing now a value from the Uniform distribution in $[0, 1]$, denoted $\mathcal{U}_{[0,1]}$, one has to invert the cdf previously computed to associate the uniformly drawn number to a specific realization y_i; this method falls under the name of *Inverse Transform Sampling*, and it is a standard approach in probability theory. In Algorithm 1 is reported the ancestral sampling for Gaussian mixtures, but the discussion above holds for any mixture model in the exponential family.

Algorithm 1: Ancestral sampling of Gaussian mixtures

```
Data: Gaussian mixture $p$ of size $n$, desired number of samples $N$.
Result: Set of samples $\boldsymbol{x}$.
$j := 0$, $\boldsymbol{x} := \{\emptyset\}$;
Compute $P(Y) : \mathbb{R}^d \rightarrow [0,1]$ as (2.14);
while $j < N$ do
    $u \sim \mathcal{U}_{[0,1]}$ ;
    $y_i := P^{-1}(u)$ ;
    $x_j \sim p(x|y_i) = \nu(x|\mu_i, \Sigma_i)$;
    $\boldsymbol{x} := \boldsymbol{x} \cup x_j$;
    j := j+1;
end
```

Maximum Likelihood for Mixtures: the Expectation-Maximization algorithm

Suppose now to address the opposite problem of sampling, that is, given a set of observations try to estimate the parameters of the corresponding mixture model. As discussed for case of single Gaussians, gammas or inverse-Wisharts, a possible approach one could think of can be the maximum likelihood estimation. Suppose a d-dimensional data set $\boldsymbol{x} = \{x_1, ..., x_N\}$ of independently drawn samples according to a Gaussian mixture distribution $p(x|\Theta)$ of size n is available. The corresponding likelihood function evaluated at $\boldsymbol{x}$ is:

$$l(\boldsymbol{x}|\Theta) \triangleq \log p(\boldsymbol{x}|\Theta) = \log p(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{j=1}^{N} \log \sum_{i=1}^{n} w_i \nu(x_j|\mu_i, \Sigma_i). \tag{2.112}$$

The first observation one can do is that now the parameters to be found involve n weighted components, hence potentially many more if compared to the single Gaussian case; moreover, the ML principle is not directly applicable, since a *logsum* term is present, hence the solution to the estimation problem is analytically intractable. Another issue arising from applying the maximum likelihood principle to Gaussian mixture models is due to the fact that it is subject to singularities and other additional problems of *identifiability*; a good corresponding discussion is provided in Chap. 9 of [3].

A solution to find maximum likelihood estimates of mixture models is the so called EM algorithm, proposed by Dempster *et al.* [9] in 1977, which is here reported for the Gaussian case:

1. Initialize the parameters of a mixture of size n and evaluate the corresponding log-likelihood over the data $\boldsymbol{x} = \{x_1, ..., x_N\}$.

2. **Expectation:** compute the matrix $R \in \mathbb{R}^{n \times N}$, containing the responsibilities for the i-th component to have generated the j-th sample, which elements are:

$$R_{i,j} = \frac{w_i \nu(x_j|\mu_i, \Sigma_i)}{\sum_{l=1}^{n} w_l \nu(x_j|\mu_l, \Sigma_l)}. \tag{2.113}$$

3. **Maximization:** Compute the *effective* number of points N_i assigned to the cluster[3] i as:

$$N_i = \sum_{j=1}^{N} R_{i,j}. \tag{2.114}$$

   Update the current mixture parameters as a *weighted* Maximum Likelihood estimation:

$$w_i^+ = \frac{N_i}{N}, \tag{2.115}$$

$$\mu_i^+ = \frac{1}{N_i} \sum_{j=1}^{N} R_{i,j} x_j, \tag{2.116}$$

$$\Sigma_i^+ = \frac{1}{N_i} \sum_{j=1}^{N} R_{i,j} (x_j - \mu_i^+)(x_j - \mu_i^+)^T. \tag{2.117}$$

4. Evaluate the log-likelihood (2.112) for the updated parameters and check if either convergence or the maximum number of allowed iterations has been reached; if not, return to step 2.

As listed above, one has to provide an initial mixture which will be refined by the EM algorithm. Nonetheless, this is a really difficult task, since a wrong initialization could push the algorithm towards singularities or inferior solutions; moreover, one has to select a model order, which, as discussed in

[3] With the term cluster is denoted a group of elements; in the EM case, it is used to refer to the Gaussian component which *encodes* a partition of the data.

2.7, can be an impactful choice. Later in this work, a possible approach to select the number of components will be discussed.

Once an initialization for the algorithm is available, the first step is to compute the responsibility matrix R, which will provide the probability of a given component to have generated a sample. Being a probability matrix, the sum by columns has to return 1. This step is known as *Expectation* and it amounts to do a *soft-clustering*[4] of the data.

After the data has been assigned "softly" to the mixture components, an update phase where the mixture parameters are re-estimated follows. This is the step where the likelihood of the current model over the data is maximized, hence the name *Maximization*.

The EM algorithm provides, hence, a sequence of refined mixture models, which will be denoted $p^{(k)} = p(\boldsymbol{x}|\Theta^{(k)})$, and it continues until a maximum number of iterations has been reached or the variation in the likelihood of the current model falls below a desired threshold.

In Algorithm 2 is reported a concise scheme of the EM algorithm.

Some additional notes

The EM is a special case of the so called Majorization-Minimization (MM) algorithm, which represents a principle more than a real algorithm. The details are omitted since a corresponding discussion is outside the scope of this work, but the MM principle is an approach to maximize or, equivalently, minimize potentially intractable cost functions by maximizing, or minimizing, a sequence of tractable local approximations; in the EM algorithm, as discussed, the complete log-likelihood can not be maximized directly, but a local tractable approximation can be exploited in each iteration. A good discussion on MM algorithms is given in [28]; regarding the EM, the corresponding detailed discussion is provide in Chap. 9 of [3].

As mentioned, the EM algorithm is not restricted to the GM case, but it can be applied to every model having latent variables. By considering gamma and inverse-Wishart mixtures, it is possible to exploit the equations reported in Sec. 2.3 to perform the equivalent maximization step. Nonetheless, the weights are now provided by the responsibilities computed in the expectation step, hence the samples are not equiprobable (weighted $\frac{1}{n}$) anymore.

[4] With the term soft-clustering the process of assigning data to clusters (represented by mixture components in this case) is done by evaluating the responsibilities, hence not providing exclusivity of the assignments.

Algorithm 2: Expectation-Maximization Algorithm for GMs

```
Data: Initial GM $p^{(0)}$ of size $n$, set of samples $\boldsymbol{x}$, maximum number
      of iterations $K$, accuracy tolerance $tol$.
Result: Refined GM of size $n$.
$k := 0$, $p^{(k)} := p^{(0)}$;
Compute $l(\boldsymbol{x}|\Theta^{(0)})$;
while $k < K$ do
    Compute $R_{i,j}$, for $i = 1, ..., n$, $j = 1, ..., N$ as in (2.111);
    Evaluate the effective number of points as in (2.114);
    Compute $w_i^+$, $\mu_i^+$, $\Sigma_i^+$ as in (2.115);
    Evaluate $l(\boldsymbol{x}|\Theta^+)$, the likelihood of the updated parameters;
    if $\|l(\boldsymbol{x}|\Theta^+) - l(\boldsymbol{x}|\Theta^{(k)})\| < tol$ then
        break;
    end
    $k := k + 1$;
    $l(\boldsymbol{x}|\Theta^{(k)}) := l(\boldsymbol{x}|\Theta^+)$;
end
```

To conclude this section, one last observation the author would like to stress out is the fact that finding a suitable number of mixture components when fitting the data is a particularly difficult problem. Selecting a very large number of components would be unjustifiable from a computational resources point of view, especially if the data can be described accurately with a much simpler model. On the other hand, providing a very simple description could be unsatisfactory from the representation accuracy point of view. As it will be discussed later in this work, if the mixture model becomes too complex (overfitting) due to the nature of the problem, it is possible to perform a *mixture reduction* in order to simplify it while preserving the accuracy. Moreover, by exploiting the optimal transport theory it is possible to obtain ”visual” information about the potential number of mixture components required to approximate the data faithfully while keeping its description lightweight.

2.9 Clustering

As mentioned in Sec. 2.8, the EM algorithm represents a *soft-clustering* algorithm which partitions of the sample data in clusters. Hence, as well as providing a mathematical tool for describing more complex uncertainties, mixture models can also be used to cluster data. In this section, a brief discussion about clustering is reported.

K-means algorithm

Let us consider the problem of identifying groups or clusters in a set of sample data. This task can be addressed both in a probabilistic way, as done for the EM, or in a non-probabilistic way, though the so called *K-means* algorithm [29]. Suppose a set of d-dimensional samples $\boldsymbol{x} = \{x_1, ..., x_N\}$ is available; the goal is to partition such data in n clusters, with n given. In order to approach this problem, it might be convenient to define n *representatives* r_i, $i = 1, ..., n$, that is d-dimensional vectors which will serve to identify the n clusters. For the j-th sample, $j = 1, ..., N$, a corresponding binary indicator variable $M_{i,j} \in \{0, 1\}$, also called *membership*, can be defined; $M_{i,j}$ will be equal to 1 if the j-th point is associated to i-th cluster, and 0 if viceversa. The associations are mutually exclusive, that is a point can be associated to a cluster only, but a cluster can contain many points.

The clustering takes place in the Euclidean space; in this regard, one can define an objective function to minimize, which amounts to the error provided by the current clustering, as follows:

$$\mathcal{J} = \sum_{i=1}^{n} \sum_{j=1}^{N} M_{i,j} \|x_j - r_i\|^2. \tag{2.118}$$

$\mathcal{J}$ is sometimes also known as *distortion measure* and it represents the sum of the squares of the distances of each sample to its representative. Once the problem has been cast in this form, the goal is to minimize (2.118) in order to obtain the smallest clustering error.

As done for the EM, or more in general from an MM perspective, such a problem can be addressed through two alternating phases, that is a phase of *assignment* and a phase of *update*. By recalling the EM scheme, the K-means algorithm can be seen as a simpler case of the said algorithm; moreover, it represents a *hard-clustering* method, since the assignments are done in an exclusive way, and not like the EM where the responsibilities take into account the assignment of a sample to several clusters. Moreover, in the EM algorithm the likelihood is being maximized, but as it will be discussed in 3.1.1, maximizing such quantity amounts to minimize the cross-entropy of the model from the data. In the K-means, there are no probabilities involved, and the sum of square distances represents the loss function to be minimized in the phase following the assignment. In this regard, it is here reported the K-means procedure:

1. Generate randomly the set $\boldsymbol{r} = \{r_i\}_{i=1}^n$ of representatives over the d-dimensional Euclidean space.

2. Perform the assignment by computing the memberships matrix $M \in \{0,1\}^{n \times N}$ which elements are:

$$M_{i,j} = \begin{cases} 1 & \text{if } i = \arg\min_k \|x_j - r_k\|^2 \\ 0 & \text{otherwise.} \end{cases} \tag{2.119}$$

3. Once the assignment has been completed, it is possible to update the representatives by at first computing the number of points associated to the i-th cluster as:

$$N_i = \sum_{j=1}^{N} M_{i,j}, \tag{2.120}$$

   and then by updating the representatives, or cluster centers, as:

$$r_i = \frac{1}{N_i} \sum_{j=1}^{N} M_{i,j} x_j, \tag{2.121}$$

   that is as the arithmetic mean of the assigned samples.

4. The algorithm continues until a maximum number of iterations has been reached, or the variation in (2.118) falls below a given tolerance.

By comparing the scheme above with the one discussed for the EM algorithm, it is possible to spot strong similarities. As for the EM, even the K-means suffers from all the discussed problems, that is it is not trivial to figure out which could be a suitable number of representatives, especially in high dimensional problems, and the resulting clustering is sensitive to the initialization. In Algorithm 3 is reported a summary of the K-means algorithm.

Algorithm 3: K-means algorithm

```
Data: Initial set of representatives $\boldsymbol{r} = \{r_i\}_{i=1}^{n}$, set of samples $\boldsymbol{x}$,
      maximum number of iterations $K$, accuracy tolerance $tol$.
Result: Refined representatives, membership matrix $M$.
$k := 0$, $\boldsymbol{r}^{(k)} := \boldsymbol{r}^{(0)}$;
Compute $\mathcal{J}^{(0)}$ as in(2.118);
while $k < K$ do
    Compute the memberships $M_{i,j}$ as in (2.119);
    Compute $N_i$ as in (2.120);
    Compute the updated representatives $\boldsymbol{r}^{+}$ as in (2.121);
    Evaluate $\mathcal{J}^{+}$;
    if $\|\mathcal{J}^{+} - \mathcal{J}^{(k)}\| < tol$ then
        break;
    end
    $k := k + 1$;
    $\mathcal{J}^{(k)} := \mathcal{J}^{+}$;
end
```

Note that the square euclidean distance is just one of the possible metrics which can be used to evaluate the inter-point distances. A more general version of the K-means algorithm is the K-medoids [3], which minimizes the following objective function:

$$\tilde{\mathcal{J}} = \sum_{i=1}^{n} \sum_{j=1}^{N} M_{i,j} \mathcal{V}(x_j, r_i). \tag{2.122}$$

where $\mathcal{V}(\cdot, \cdot)$ is a generic distance measure between points in the euclidean space. Moreover, the K-medoids can be further generalized if one considers a different space, e.g., the space of probability distributions. In this regard, in the next chapter, such a generalization will be discussed in the context of mixture refinement.

2.10 Bayesian Filtering

In order to further stress the importance of the problem that will be addressed in this work, namely the Mixture Reduction problem, the author decided to report a discussion about the target tracking in presence of clutter, another application where mixture models are broadly used and, moreover, where the exponential growth in the number of components appears, hence requiring approximations. Nonetheless, target tracking is a very broad topic, and providing an exhaustive discussion would be cumbersome, hence only the main aspects will be treated; in this regard, good textbooks one can use as reference are [4, 11].

2.10.1 The Filtering Problem

The filtering problem is concerned with establishing the best estimate for the true value of some quantity from an incomplete, potentially noisy set of observations done on that quantity. The Bayesian inference framework represents an optimal solution to this kind of problems (see 2.5).

Bayesian inference for dynamical systems

In the Bayesian framework, the optimal filtering problem is considered to be a *statistical inversion* problem, where the unknown quantity is a vector valued time series $\{s_0, s_1, ...\}$ which is observed through a set of noisy measurements $\{z_1, z_2, ...\}$ as reported in Fig. 2.4

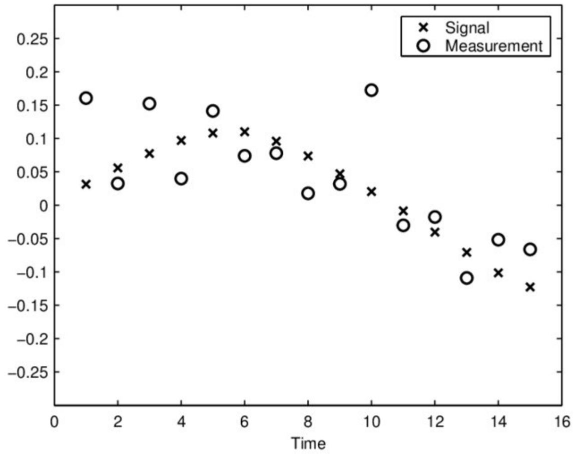

Figure 2.4: Representation of a time series observed through a set of noisy measurements [1].

For the sake of discussion, only the discrete time case will be addressed.

The purpose of statistical inversion is to estimate the hidden states $s_{0:T} = \{s_0, s_1, ..., s_T\}$ from the observations $z_{1:T} = \{z_1, z_2, ..., z_T\}$, which is equivalent to compute the joint posterior distribution by applying the Bayes' rule (2.21) as:

$$p(s_{0:T}|z_{1:T}) = \frac{p(z_{1:T}|s_{0:T})p(s_{0:T})}{p(z_{1:T})}. \tag{2.123}$$

Such a formulation, though, has clear disadvantages, that is for any new measurement, the full posterior distribution has to be recomputed; moreover, the posterior dimensionality increases at each step, hence the statistical inversion becomes computationally intractable quickly. Nonetheless, in many filtering problems one is interested only in the state estimate at the current time, hence the full posterior is not necessary.

This hindrance can be solved by assuming that the considered dynamical systems are *Markov sequences*, which implies that given the state s_k, only the observation z_k depends on such state (conditional independence), hence (2.123) can be rewritten as:

$$p(s_k|z_{1:k}) = \frac{p(z_k|s_k)p(s_k|z_{1:k-1})}{p(z_k|z_{1:k-1})} = \frac{p(z_k|s_k)p(s_k|z_{1:k-1})}{\int p(z_k|s_k)p(s_k|z_{1:k-1})\mathrm{d}s_k}. \tag{2.124}$$

Moreover, the Markov assumption allows to break the full posterior distribution estimation in a recursive computation, hence by propagating over time the knowledge one has about the dynamical system state and by incorporating at each recursion step the information deriving from a new observation. By recalling the argumentation done in 2.5, it is straightforward to recognize the involved quantities in the above recursion. $p(z_k|s_k)$ is the likelihood function, $p(s_k|z_{1:k-1})$ is the prior distribution, while $p(z_k|z_{1:k-1})$ is the evidence, or marginal likelihood function. In the context of dynamical systems those quantities can be further specialized: $p(s_k|z_{1:k-1})$ is called *state prediction density*, $p(z_k|z_{1:k-1})$ can be also termed as *predicted likelihood*, while $p(z_k|s_k)$ is generally called *measurement model*, since it describes how a noisy sensor observes the true state of the system.

By exploiting (2.18), let us expand the prior distribution in (2.124) as follows:

$$p(s_k|z_{1:k-1}) = \int p(s_k|s_{k-1})p(s_{k-1}|z_{1:k-1})\mathrm{d}s_{k-1}. \tag{2.125}$$

The one above is known as the *Chapman-Kolmogorov* equation, which allows to predict the posterior distribution from the previous step to the current one.

$p(s_k|s_{k-1})$ is called *motion model* and describes how the dynamical system evolves over time. Hence, given an initial distribution $p(s_0|z_0) = p(s_0)$ (since the first measurement is obtained at time step 1), it is possible to implement a *prediction-correction* scheme by at first predicting $p(s_1|z_0)$ by means of (2.125), and then by obtaining the updated posterior distribution $p(s_1|z_1)$ by means of the Bayes' formula (2.124). From a practical point of view, one does not have any information coming from observations at time 0; hence, it would be more correct to assume to have an initial guess $p(s_1)$ which is a totally *subjective* guess of the system state, and by updating it as the first observation arrives instead of performing a prediction. Thus said, a practical recursion could be something like:

- Initial guess $p(s_1)$.
- Given z_1, perform the update (2.124) to obtain $p(s_1|z_1)$.
- Perform the prediction (2.125) to obtain $p(s_2|z_1)$.
- Given z_2, perform the update (2.124) to obtain $p(s_2|z_2)$,
- ... and so on.

Motion and Measurement models

As mentioned, the motion model describes how the system evolves over time. From a probabilistic point of view, it is the *state transition density.* There exist many commonly used motion models in the literature, and they can be linear or nonlinear. For instance, in the class of linear motion models one finds the Constant Velocity (CV) model, which assumes a first order motion of an object, the Constant Acceleration (CA) model, which assumes a second order motion, while for the nonlinear case an example can be the Coordinated Turn (CT) model. Further details are omitted, but, again, good references on these topics can be [4, 11].

The measurement model describes instead the relation between the state of the dynamical system and the corresponding observation; such a model is *sensor-dependent*, since even several sensors in the same class (e.g., lidars, radars, IMUs etc.) can have different technical features, hence different ways to observe a quantity of interest. Moreover, given a sensor, different measurement models may be required to observe different targets (e.g., a radar observing both a racecar and a bicycle).

Bayesian inference solution for dynamical systems

The Bayesian recursion (2.124) does not have in general a closed form solution. Nonetheless, for the linear-Gaussian dynamical systems the Kalman Filter (KF) [7] provides the optimal corresponding solution in closed-form. For nonlinear-non-Gaussian dynamical systems there are no optimal closed form solutions, but many approximations such as the Extended Kalman Filter (EKF) [1], Unscented Kalman Filter (UKF) [30], Cubature Kalman Filter (CKF) [31], Particle Filter (PF) [32] and many more have been provided.

The Kalman Filter

For Additive White Gaussian Noise (AWGN) linear models, it has been said that the KF is a solution to the Bayesian inference. Let us consider the following model:

$$\begin{cases} s_k = F_{k-1}s_{k-1} + \xi_{k-1}, \\ z_k = H_k s_k + r_k, \end{cases} \tag{2.126}$$

where:

- $s_k \in \mathbb{R}^d$ is the state vector,
- $F_{k-1} \in \mathbb{R}^{d\times d}$ is the state transition matrix at one step,
- $\xi_{k-1} \in \mathbb{R}^d$ is the process noise vector, with $\xi_{k-1} \sim \nu(\cdot|\mathbf{0}, Q_{k-1})$,
- $z_k \in \mathbb{R}^m$ is the observation on the system,
- $H_k \in \mathbb{R}^{m\times d}$ represents the relationship between the state and the measurement at time k,
- $r_k \in \mathbb{R}^m$ is the measurement noise vector, with $r_k \sim \nu(\cdot|\mathbf{0}, R_k)$.

In the case of the previous system, one has:

- Initial prior: $p(s_0) = \nu(s_0|\hat{s}_0, P_0)$,
- Posterior density: $p(s_k|z_{1:k}) = \nu(s_k|\hat{s}_{k|k}, P_{k|k})$,
- Motion model: $p(s_k|s_{k-1}) = \nu(s_k|F_{k-1}s_{k-1}, Q_{k-1})$,
- Measurement model: $p(z_k|s_k) = \nu(z_k|H_k s_k, R_k)$

where:

- $\hat{s}_{k|k}$ is the expected value of the posterior density,
- $P_{k|k}$ is the covariance matrix (representing uncertainty) of the posterior density,
- the subscript $k|k$, or $k|k-1$, denotes respectively the distribution incorporating the measurements up to the current step k, and the distribution (predicted) incorporating the information up to the previous time step $k-1$. Such notation will help to unclutter the notation in the equations which will follow.

In general, only the state estimate $\hat{s}_{k|k}$ (e.g. the position of the car or the orientation of an airplane) is used for practical purposes.
The conditional mean $\hat{s}_{k|k} = \mathbb{E}(s_k|z_{1:k})$ is a minimizer, that is the best estimate one can achieve for the state given the observations. Following are reported all the necessary equations to implement a KF recursion:

$$\begin{aligned}
&Prediction: \begin{cases} \hat{s}_{k|k-1} = F_{k-1}\hat{s}_{k-1|k-1}, & \text{Use model to predict} \\ P_{k|k-1} = F_{k-1}P_{k-1|k-1}F_{k-1}^T + Q_{k-1}, & \text{Increase uncertainty} \end{cases} \\
&Update: \begin{cases} \hat{z}_k = H_k\hat{s}_{k|k-1}, & \text{Predicted measurement} \\ \epsilon_k = z_k - \hat{z}_k, & \text{Innovation} \\ S_k = H_kP_{k|k-1}H_k^T + R_k, & \text{Innovation covariance} \\ K_k = P_{k|k-1}H_k^TS_k^{-1}, & \text{Kalman gain} \\ \hat{s}_{k|k} = \hat{s}_{k|k-1} + K_k\epsilon_k, & \text{State update, weighted average} \\ P_{k|k} = P_{k|k-1} - K_kH_kP_{k|k-1}, & \text{Decrease uncertainty} \end{cases}
\end{aligned} \tag{2.127}$$

Another quantity of interest worth to be reported is the following:

$$Predicted\ Likelihood: p(z_k|z_{1:k-1}) = \nu(z_k|\hat{z}_k, S_k) \tag{2.128}$$

Such a quantity can be used to perform statistical tests on the filter performances; corresponding details can be found [11].

Single Object Tracking in clutter

In the previous section it has been implicitly assumed that one, and only one, observation is acquired at each time step; moreover, it was generated over the quantity of interest. Is this always true? When sensors like cameras, lidars or radars are used, the answer to this question is no.

In general, the filtering problem becomes more challenging, since:

- *Misdetections* can occur, that is the object is present in the sensor field of view, but it is not detected.
- *Clutter* can be present, that is several observations at each time step are acquired, and they do not belong only to the object of interest. Such observations may be generated from reflections, sensor malfunctioning or other entities which are not of interest.
- *Unknown data association*, that is it is not known which measurement belongs (if any) to the object of interest. In order to obtain an optimal estimate of the system state, all the possible associations between the measurements and the previous posterior distribution estimate have to be taken into account.

Detections and Misdetections

How can one model the chance of misdetection for an object? A common approach is to say that the object is detected with probability $P^D(s_k)$ and it generates a measurement according to the measurement model $p(o_k|s_k)$, where o_k is the observation associated to the object.
Let us define the set of observations for the object as the following matrix:

$$O_k = \begin{cases} [\,], & \text{if the object is undetected,} \\ o_k, & \text{otherwise.} \end{cases} \tag{2.129}$$

With $n(O_k)$ will be denoted the cardinality, or number of columns, of O_k. From the previous definitions, it follows that given s_k, $n(O_k)$ is Bernoulli (see (2.51)) distributed:

$$n(O_k) = \begin{cases} 1 & \text{with probability } P^D(s_k), \\ 0 & \text{with probability } 1 - P^D(s_k), \end{cases} \tag{2.130}$$

from which follows that:

$$p(O_k|s_k) = \begin{cases} 1 - P^D(s_k) & \text{if } O_k = [\,], \\ P^D(s_k)p(o_k|s_k) & \text{if } O_k = o_k. \end{cases} \tag{2.131}$$

This form for the likelihood function captures both the probability of detection and, if detected, the distribution of the detection according to the measurement model.

Clutter

As mentioned, in the field of view of a sensor can be present other unwanted entities which may be responsible for the generation of additional observations in the same time step. Let us define the complete observation matrix as:

$$Z_k = s(O_k, C_k), \tag{2.132}$$

where $s(\cdot)$ is an operator which randomly shuffles column vectors; this will model the unknown origin of the measurements. How can the clutter be modelled in a probabilistic perspective? In this regard, a stochastic model is required for both:

- number of clutter detections $n(C_k)$,
- vectors in C_k.

If one considers the volume V of a sensor field of view and λ to be the expected number of clutter detections per unit volume, then the clutter can be modelled as a Poisson Point Process (PPP).

Clutter Poisson Point Process

The Poisson point process is the default model for clutter $C_k = [c_k^1, ..., c_k^{m_k^c}]$, with $m_k^c \sim \pi(x|\lambda V)$ (see (2.53)) number of clutter-originated measurements. Moreover, it holds that given m_k^c, the vectors $c_k^1, ..., c_k^{m_k^c}$ are i.i.d. with $c_k^i \sim \mathcal{U}_V$, where V is the sensor field of view, and $\mathcal{U}_V$ is the uniform distribution over such volume. In general, the PPPs are parameterized using either:

- an intensity function, $\lambda_c(c) > 0$, or
- a combination of:

$$\begin{cases} \bar{\lambda}_c = \int \lambda_c(c) \mathrm{d}c & \text{rate,} \\ f_c(c) = \frac{\lambda_c(c)}{\bar{\lambda}_c} & \text{spatial pdf.} \end{cases}$$

The PPP distribution is of the form:

$$p(C_k) = \frac{e^{-\bar{\lambda}_c}}{m_k^c!} \prod_{i=1}^{m_k^c} \lambda_c(c_k^i). \tag{2.133}$$

Data Association

As discussed, the presence of clutter and misdetections introduces the requirement of considering all the possible associations between the current prior and the available measurements in order to compute the posterior distribution in a Bayesian inference context. In this regard, it is necessary to formalize the concept of Data Association (DA). Let us define the following quantity:

$$\tau = \begin{cases} i > 0, & \text{if } z^i \text{ is an object detection,} \\ 0, & \text{if the object is undetected.} \end{cases} \tag{2.134}$$

Here follows an example to better understand such notation. Suppose at a given time the measurement matrix $Z = [z^1, z^2, z^3]$ is obtained; if $\tau = 2$, then one is saying that z^2 is an object detection and z^1 and z^3 are clutter. If, instead, $\tau = 0$, then the object has been misdetected and all the observations are clutter.

Posterior distributions in Single-Object Tracking

Given the previously defined mathematical models for clutter, misdetections and data associations, it is possible to write the full measurement model as:

$$\begin{aligned} p(Z_k|s_k) = & \left[(1 - P^D(s_k)) + P^D(s_k) \sum_{\tau_k=1}^{m_k} \frac{p(z_k^{\tau_k}|s_k)}{\lambda_c(z_k^{\tau_k})}\right] \frac{e^{-\bar{\lambda}_c}}{m_k!} \prod_{i=1}^{m_k} \lambda_c(z_k^i) = \\ & = \sum_{\tau_k=0}^{m_k} p(Z_k, m_k, \tau_k|s_k), \end{aligned} \tag{2.135}$$

where m_k is the number of measurement vectors in Z_k at time k and τ_k is the DA variable.

Note: Each τ represents a state *hypothesis* associated to a given observation; moreover, the posterior density $p(s_k|Z_{1:k})$ at time k will contain all the different hypotheses from all the time steps. In this regard, it is possible to write posterior distribution, by exploiting (2.17), as:

$$p(s_k|Z_{1:k}) = \sum_{\tau_{1:k}} p(s_k|\tau_{1:k}, Z_{1:k}) p(\tau_{1:k}|Z_{1:k}), \tag{2.136}$$

where $\sum_{\tau_{1:k}} = \sum_{\tau_1=0}^{m_1} \sum_{\tau_2=0}^{m_2} \cdots \sum_{\tau_k=0}^{m_k}$, while the prior can be rewritten as:

$$p(s_k|Z_{1:k-1}) = \sum_{\tau_{1:k-1}} p(s_k|\tau_{1:k-1}, Z_{1:k-1}) p(\tau_{1:k-1}|Z_{1:k-1}). \tag{2.137}$$

From the previous equations it is possible to figure out that:

- At each time step k, there are $m_k + 1$ new data association hypotheses.
- The number of possible association sequences at time k is:

  $$\prod_{i=1}^{k} (m_i + 1) = (m_1 + 1) \times \cdots \times (m_k + 1), \tag{2.138}$$

  that is the number of hypotheses for the system state grows incredibly fast with k.

By plugging together all the parts, it can be proven that the resulting posterior distribution will have the following form:

$$p(s_k|Z_{1:k}) = \sum_{\tau_{1:k}} w^{\tau_{1:k}} q_{k|k}^{\tau_{1:k}}(s_k), \tag{2.139}$$

that is the posterior distribution will be a mixture of densities[5] (see 2.8, where $\boldsymbol{w}^{\tau_{1:k}}$ is a pmf describing the probability of a given association sequence, and $\boldsymbol{q}^{\tau_{1:k}}$ will represent the hypotheses at time step k.

Prediction Equations

Let us assume that a posterior density from the previous time step is available in the form:

$$p(s_{k-1}|Z_{1:k-1}) = \sum_{\tau_{1:k-1}} w^{\tau_{1:k-1}} q_{k-1|k-1}^{\tau_{1:k-1}}(s_{k-1}), \tag{2.140}$$

[5] some small abuse of notation has been made, but the sum will be over all the existing hypotheses at time k, which are expontially growing over time.

then, by applying (2.125), one obtains:

$$\begin{aligned} p(s_k|Z_{1:k-1}) &= \int p(s_{k-1}|Z_{1:k-1})p(s_k|s_{k-1})\mathrm{d}s_{k-1} \\ &= \sum_{\tau_{1:k-1}} w^{\tau_{1:k-1}} \underbrace{\int q_{k-1|k-1}^{\tau_{1:k-1}}(s_{k-1})p(s_k|s_{k-1})\mathrm{d}s_{k-1}}_{\triangleq q_{k|k-1}^{\tau_{1:k-1}}(s_k)} \\ &= \sum_{\tau_{1:k-1}} w^{\tau_{1:k-1}} q_{k|k-1}^{\tau_{1:k-1}}(s_k). \end{aligned} \tag{2.141}$$

Note: the weights remain unchanged; the standard prediction is performed for each hypothesis.

Update Equations

By recalling (2.135), it is possible to write:

$$\begin{aligned} p(s_k|Z_{1:k}) \propto &\sum_{\tau_{1:k-1}} w^{\tau_{1:k-1}} q_{k|k-1}^{\tau_{1:k-1}}(s_k)(1-P^D(s_k)) \\ &+ \sum_{\tau_{1:k-1}} \sum_{\tau_k=1}^{m_k} \frac{1}{\lambda_c(z_k^{\tau_k})} w^{\tau_{1:k-1}} q_{k|k-1}^{\tau_{1:k-1}}(s_k)P^D(s_k)p(z_k^{\tau_k}|s_k). \end{aligned} \tag{2.142}$$

For every pair of hypotheses $(\tau_{1:k-1}, \tau_k)$ a new hypothesis is generated, which is indexed as $\tau_{1:k}$.

Let us denote with $\tilde{w}^{\tau_{1:k}}$ the unnormalized weight associated to a sequence of associations $\tau_{1:k}$; after the update step, the posterior distribution will be $p(s_k|Z_{1:k}) = \sum_{\tau_{1:k}} w^{\tau_{1:k}} q_{k|k}^{\tau_{1:k}}(s_k)$, where $w^{\tau_{1:k}} \propto \tilde{w}^{\tau_{1:k}}$ and:

$$\begin{aligned} &\tau_k = 0 \text{ (misdetection)} : \begin{cases} \tilde{w}^{\tau_{1:k}} = w^{\tau_{1:k-1}} \int q_{k|k-1}^{\tau_{1:k-1}}(s_k)(1-P^D(s_k))\mathrm{d}s_k, \\ q_{k|k}^{\tau_{1:k}}(s_k) \propto q_{k|k-1}^{\tau_{1:k-1}}(s_k)(1-P^D(s_k)), \end{cases} \\ \\ &\tau_k > 0 \text{ (obj. detection)} : \begin{cases} \tilde{w}^{\tau_{1:k}} = \frac{w^{\tau_{1:k-1}}}{\lambda_c(z_k^{\tau_k})} \int q_{k|k-1}^{\tau_{1:k-1}}(s_k)P^D(s_k)p(z_k^{\tau_k}|s_k)\mathrm{d}s_k, \\ q_{k|k}^{\tau_{1:k}}(s_k) \propto q_{k|k-1}^{\tau_{1:k-1}}(s_k)P^D(s_k)p(z_k^{\tau}|s_k). \end{cases} \end{aligned} \tag{2.143}$$

To conclude this section, the author wants to remark that, for the problem of target tracking in clutter, the posterior distribution obtained by the optimal

Bayesian recursions yields a mixture of densities which grows exponentially in the number of components over time. If no approximations are introduced, the uncertainty representation about the system state becomes computationally intractable after few steps. There exists several algorithms as the Nearest Neighbour (NN) or Probabilistic Data Association (PDA) filters [11], which collapse all the hypotheses into a single one by means of moment-matching, discussed in the next chapter. Nonetheless, approximating a multimodal distribution with a single density can be a rather crude approximation, which has proven to often lead to filter divergence. Alternatively, as discussed in [23] for the Gaussian Sum (GS) filter, it is possible to propagate over time a mixture of densities which has a controlled number of hypotheses; as it will be discussed in Chapter 3, there have been proposed many approaches to control the number of hypotheses in a mixture, but given the difficulty of the problem, heuristics favoring ease of computations have always been adopted, without investigating the actual nature of many algorithms. The discussion here reported is only regarding the single-object case; when multiple objects are considered, the number of hypotheses grows even faster [13]. To address the Multi-Object Tracking (MOT) problem, a new and elegant framework based on RFSs, namely the Finite Sets Statistics framework [8], has been proposed by R. Mahler, which laid the foundations for many algorithms like the Gaussian Mixture PHD (GMPHD) filter [25], or the Poisson Multi-Bernoulli Mixture (PMBM) filter [33]. Those algorithms work with intensities rather than mixtures, but the problem of combinatorial explosion in the number of components still persists. At this point, it should be clear why the MRP represents an important problem, since the vast majority of filtering/tracking algorithms today used suffer from the problem of hypotheses management. Moreover, even in problems of unsupervised learning, as briefly mentioned in Sec. 2.8, or many others not reported here (Belief-Propagation, Kernel-Density Estimation, etc.), the mixture reduction problem arises. In Chapter 3, such a problem will be formalized and existing approaches will be discussed. In Chapter 4 a new framework will be proposed to address this problem in a consistent and efficient way.

Chapter 3

The Mixture Reduction Problem

3.1 Dissimilarity Measures

An important concept when addressing the Mixture Reduction Problem is that of *dissimilarity*. The dissimilarity, also called deviation, distortion or divergence, between two probability distributions is a measure (D-measure for short) of how dissimilar they are. In general, for all pairs of distributions p and q over $\mathbb{R}^d$, a D-measure satisfies the following properties:

$$D(p\|q) \geq 0, \qquad \text{(nonnegativity);} \tag{3.1}$$

$$D(p\|q) = 0 \iff p = q, \qquad \text{(identity of indiscernibles).} \tag{3.2}$$

If, for any triple p, q, h, also the following hold true

$$D(p\|q) = D(q\|p), \qquad \text{(symmetry);} \tag{3.3}$$

$$D(p\|q) \leq D(p\|h) + D(h\|q), \qquad \text{(triangle inequality),} \tag{3.4}$$

then the dissimilarity $D(\cdot\|\cdot)$ is a *distance* and defines a metric in the space of distributions.

In the literature there exists a broad range of D-measures, and they all exhibit different features, both in terms of analytical properties and peculiarities. In this regard, a preliminary discussion about the features of several D-measures here presented will be provided in section 3.3, and later specialized in section 5.1.

As already mentioned, all the discussions which will follow are quite general and can be applied to any distribution in the exponential family. Nonetheless, reporting formulas for all the classes in such family would be burdensome and not necessary for the goals of this work; in this regard, the discussion will restrict to the Gaussian case and, when required, additional formulas for other classes will be provided.

Note: if the generic Gaussian density ν is considered, then the dissimilarity between a Gaussian pdf ν_i and ν will be denoted $D(\nu_i \| \nu)$ (or equivalently $D(\nu \| \nu_i)$ if D is not symmetric, and the dissimilarity of ν_i from ν is sought).

Note that closed formulas are usually available when considering densities in the exponential family, but, if mixtures built accordingly are considered, only in very few cases one can evaluate analytically the corresponding dissimilarity. As it will be discussed in Chapter 4, evaluating pairwise dissimilarities between densities is sufficient to induce dissimilarities between mixtures.

In Appendix A.1 all the D-measure here discussed will be reported together with some useful formulae and identities; to ease the discussion, though, all the necessary quantities will be defined in the process.

Some additional notes on D-measures

A D-measure as defined previously should also preserve the ordering in the space of distributions. Let us consider a parallelism with the euclidean space. Given a point x in such space, any other point y different from x will have an euclidean distance $d(x, y) > 0$. Moreover, given a line starting from x and ending in y, and by moving along the direction from x to y, all the points lying on it will be increasingly distant from x and closer to y, that is there is an *ordering* in the euclidean space if the euclidean distance is considered. If one considers the space of distributions and a reference distribution $p(x)$ over that space, then any other $q(x) \neq p(x)$ will have a greater than zero dissimilarity from $p(x)$, that is (3.1) and (3.2) hold. Imagine now that one can define a "line" in the space of distributions which starts from $p(x)$ and ends in $q(x)$; then, a dissimilarity measure as such should preserve the ordering as discussed for the euclidean case, that is all the distributions falling on the line joining $p(x)$ to $q(x)$ should have a non-decreasing dissimilarity from $p(x)$ and a non-increasing dissimilarity from $q(x)$. Thus said, most of the D-measures considered in this work satisfy such a property, but the author preferred to remark this argument for completeness.

3.1.1 Kullback-Leibler Divergence

The gold standard in terms of statistical divergences is the so called Kullback-Leibler Mean Information (KLI) [34], which, given two distributions p and q over $\mathbb{R}^d$, is defined as follows:

$$D_{KL}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x. \tag{3.5}$$

Such a D-measure satisfies properties (3.1) and (3.2), so it is a *directed* divergence, since it does not possess symmetry. In the original work [34], the term divergence is used for the symmetrized form of the above equation, which will be presented in Sec. 3.1.2. Nonetheless, in the literature the term Kullback-Leibler Divergence (KLD) is often used to denote (3.5); for the remainder of this paper such naming convention will be adopted.

The KLD, also known as *differential relative entropy*, can be seen as the amount of *information* lost if p is approximated by q. More in detail, it can be seen as the *expected log-likelihood ratio* of p and q, which can be denoted as $\mathbb{E}_p[\log\left(\frac{p}{q}\right)]$. By exploiting the logarithm properties, one obtains:

$$D_{KL}(p\|q) = \int p(x) \log p(x) \mathrm{d}x - \int p(x) \log q(x) \mathrm{d}x = -H[p] + H_\times[p,q]. \tag{3.6}$$

where $H[p]$ is the differential entropy (2.92) of p and $H_\times[p,q]$ is the differential cross-entropy (2.98) between p and q, as defined in 2.6.

Maximum Likelihood estimation and KLD

Let us consider the parametric distribution $q(x|\theta) \in \mathcal{Q}$ that is as close as possible to a distribution p in terms of D_{KL}, that is:

$$q^* = \arg\min_\theta D_{KL}(p\|q(\cdot|\theta)) = \arg\min_\theta \int p(x) \log p(x) \mathrm{d}x - \int p(x) \log q(x|\theta) \mathrm{d}x. \tag{3.7}$$

Note that the entropy of p part of the integral does not depend on θ, so it can be considered as a constant in the above problem.

Suppose now that p is unknown, but it is possible to sample from it; one can then build the corresponding *empirical distribution* as:

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i), \tag{3.8}$$

where $\{x_i\}_{i=1}^N$ are N i.i.d. samples drawn from p and δ is the *Dirac's delta*. Using the properties of the delta one obtains:

$$D_{KL}(\hat{p}\|q(\cdot|\theta)) = - H[\hat{p}] - \int \hat{p}(x) \log q(x|\theta) \mathrm{d}x = \tag{3.9}$$

$$= - H[\hat{p}] - \int \left[\frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i) \right] \log q(x|\theta) \mathrm{d}x = \tag{3.10}$$

$$= - H[\hat{p}] - \frac{1}{N} \sum_{i=1}^{N} \log q(x_i|\theta). \tag{3.11}$$

Minimizing the quantity above w.r.t. to θ, that is by minimizing the cross-entropy term, is equivalent to maximize the average negative log likelihood of the parameter θ of q on the samples drawn from p; hence, it is possible to say that minimizing the D_{KL} to the empirical distribution is equivalent to find the maximum likelihood estimate of the model q [16].

As discussed, the KLD does not possess the symmetry property and the order in which the distributions are considered can be impactful in terms of the outcome. In this regard, the Forward Kullback-Leibler Divergence (FKLD), $D_{FKL}(p\|q)$ for short, is the one defined as in (3.5), while the Reverse Kullback-Leibler Divergence (RKLD), $D_{RKL}(p\|q)$ for short, denotes the divergence obtained by swapping the order of the arguments, that is $D_{RKL}(p\|q) = D_{FKL}(q\|p)$.
If two Gaussian densities $\nu_i(x)$ and $\nu_j(x)$ are considered, then one obtains:

$$D_{FKL}(\nu_i\|\nu_j) = \frac{1}{2}(\mathrm{tr}(\Sigma_j^{-1}\Sigma_i) + (\mu_i - \mu_j)^T \Sigma_j^{-1}(\mu_i - \mu_j) - d + \log \frac{|\Sigma_j|}{|\Sigma_i|}). \tag{3.12}$$

The $D_{RKL}(\nu_i\|\nu_j)$ can be obtained by swapping the means and the covariances in equation (3.12).
In the gamma case, given two distributions $\gamma_i(\chi)$ and $\gamma_j(\chi)$, the corresponding D_{FKL} is:

$$D_{FKL}(\gamma_i\|\gamma_j) = \kappa_j \log \frac{\omega_j}{\omega_i} + \log \frac{\Gamma(\kappa_j)}{\Gamma(\kappa_i)} + (\kappa_i - \kappa_j)\psi_0(\kappa_i) + \kappa_i \frac{\omega_i - \omega_j}{\omega_j}. \tag{3.13}$$

In the inverse-Wishart case, given two distributions $\varphi_i(\mathcal{Y})$ and $\varphi_j(\mathcal{Y})$, one obtains:

$$D_{FKL}(\varphi_i\|\varphi_j) = \log \frac{\Gamma_d\left(\frac{v_j-d-1}{2}\right)}{\Gamma_d\left(\frac{v_i-d-1}{2}\right)} + \left(\frac{v_i-d-1}{2}\right)\mathrm{tr}\big(V_i^{-1}V_j\big) - \left(\frac{v_i-d-1}{2}\right)d+ \\ -\left(\frac{v_j-d-1}{2}\right)\log|V_i^{-1}V_j| - \frac{v_j-v_i}{2}\sum_{m=1}^{d}\psi_0\left(\frac{v_i-d-m}{2}\right). \tag{3.14}$$

When applied to mixtures, the D_{KL} does not admit a closed form; this may represent a significant drawback for a D-measure when employed in the MRP, since one has to resort to tractable approximations in order to evaluate the dissimilarity between mixtures. In any case, investigating such approximations could provide useful insights on the kind of usage one can make of them; more on this will be discussed in Chapter 4.

3.1.2 Skew Jeffreys' Divergence

The Jeffreys' Divergence is a symmetrization of the KLD (more properly of the KLI). Its skew version depends on a parameter $\alpha \in [0, 1]$ and, given two distributions p and q over $\mathbb{R}^d$, is defined as follows:

$$D_J^\alpha(p\|q) = (1-\alpha)D_{FKL}(p\|q) + \alpha D_{RKL}(p\|q), \tag{3.15}$$

Symmetrized Kullback-Leibler Divergence

For $\alpha = 0.5$ is obtained the so called Symmetrized Kullback-Leibler Divergence (SKLD) (or Jeffrey's divergence), D_{SKL} for short, defined as:

$$D_{SKL}(p\|q) = \frac{1}{2}\big(D_{FKL}(p\|q) + D_{RKL}(p\|q)\big) = \frac{1}{2}\int (p(x)-q(x))\log\frac{p(x)}{q(x)}\mathrm{d}x. \tag{3.16}$$

Such D-measure possesses properties (3.1), (3.2) and (3.3). By plugging (3.12) in (3.15) and (3.16) is obtained the corresponding closed form for the Gaussian case:

$$D_{SKL}(\nu_i\|\nu_j) = \frac{1}{4}\big[\mathrm{tr}(\Sigma_j^{-1}\Sigma_i + \Sigma_i^{-1}\Sigma_j) + (\mu_i-\mu_j)^T(\Sigma_i^{-1}+\Sigma_j^{-1})(\mu_i-\mu_j) - 2d\big]. \tag{3.17}$$

As for the D_{KL}, the D_{SKL} does not admit a closed form when applied to mixtures.

3.1.3 Squared 2-Wasserstein Distance

The Squared 2-Wasserstein (W2), D_{W2} for short, between two distributions p and q over $\mathbb{R}^d$ is defined as follows:

$$D_{W2}(p\|q) = \inf_{\pi} \mathbb{E}\{\|X-Y\|_2^2\} = \inf_{\pi} \iint \|x-y\|_2^2 \pi(x,y)\mathrm{d}x\mathrm{d}y, \tag{3.18}$$

where $X \sim p(x)$, $Y \sim q(y)$, $\mathbb{E}(\cdot)$ is the expectation, and $\pi(x,y)$ is any joint distribution of (X,Y) that has $p(x)$ and $q(y)$ as marginals. This D-measure satisfies (3.1), (3.2) and (3.3). The D_{W2} between two Gaussians $\nu_i(x)$ and $\nu_j(x)$ is:

$$D_{W2}(\nu_i\|\nu_j) = \|\mu_i - \mu_j\|_2^2 + \mathrm{tr}\big(\Sigma_i + \Sigma_j - 2\big(\Sigma_i^{\frac{1}{2}}\Sigma_j\Sigma_i^{\frac{1}{2}}\big)^{\frac{1}{2}}\big). \tag{3.19}$$

If mixtures are considered, the D_{W2} does not admit a closed form.

3.1.4 Likeness-based Dissimilarity Measures

Rather than a single D-measure, in the following is defined a whole class under the name of Likeness-Based (LB) family [35]. Among the reported D-measures, the set here considered possesses closed forms in the mixture case. Given two distributions p and q over $\mathbb{R}^d$, let us define the following quantities:

$$J^{p,p} = \int p(x)^2 \mathrm{d}x, \tag{3.20}$$

$$J^{p,q} = \int p(x)q(x) \mathrm{d}x, \tag{3.21}$$

$$J^{q,q} = \int q(x)^2 \mathrm{d}x. \tag{3.22}$$

In the literature those terms are known as Cross Information Potentials (CIPs) [36] or *Likenesses* [35, 37]. For the remainder of this work the author will adopt the latter naming convention.

If two Gaussians ν_i and ν_j are considered, one obtains:

$$J^{i,i} = \nu(\mu_i|\mu_i, 2\Sigma_i) \underbrace{\int \nu(x|\bar{\mu}_{i,i}, \bar{\Sigma}_{i,i})\mathrm{d}x}_{=1} = |4\pi\Sigma_i|^{-\frac{1}{2}}, \tag{3.23}$$

$$J^{i,j} = \nu(\mu_i|\mu_j, \Sigma_i + \Sigma_j) \underbrace{\int \nu(x|\bar{\mu}_{i,j}, \bar{\Sigma}_{i,j})\mathrm{d}x,}_{=1} \tag{3.24}$$

$$J^{j,j} = \nu(\mu_j|\mu_j, 2\Sigma_j) \underbrace{\int \nu(x|\bar{\mu}_{j,j}, \bar{\Sigma}_{j,j})\mathrm{d}x}_{=1} = |4\pi\Sigma_j|^{-\frac{1}{2}}, \tag{3.25}$$

where $J^{i,i}$ and $J^{j,j}$ are respectively the *self-likenesses* of ν_i and ν_j, while $J^{i,j}$ is the *cross-likeness* between ν_i and ν_j; the indices i and j are a compact notation for ν_i and ν_j, and:

$$\bar{\mu}_{i,j} = \big(\Sigma_i^{-1} + \Sigma_j^{-1}\big)^{-1}\big(\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j\big), \tag{3.26}$$

$$\bar{\Sigma}_{i,j} = \big(\Sigma_i^{-1} + \Sigma_j^{-1}\big)^{-1}. \tag{3.27}$$

When a generic Gaussian density $\nu(x|\mu, \Sigma)$ is considered, the notation is changed to $J^{i,\nu}$ and $J^{\nu,\nu}$, which denote respectively the cross-likeness between the i-th Gaussian pdf and a generic Gaussian component ν, and the self-likeness of a generic Gaussian component ν.
If two Gaussian mixtures $p^a = p(x|\Theta^a) = (\boldsymbol{w}^a)^T\boldsymbol{\nu}^a(x|\boldsymbol{\theta}^a) = \sum_{i=1}^{n^a} w_i^a\nu_i^a$ and $p^b(x|\Theta^b) = (\boldsymbol{w}^b)^T\boldsymbol{\nu}^b(x|\boldsymbol{\theta}^b) = \sum_{j=1}^{n^b} w_j^b\nu_j^b$ are considered, then one obtains:

$$J^{a,a}(\Theta^a) = \int (\boldsymbol{w}^a)^T\boldsymbol{\nu}^a(x|\boldsymbol{\theta}^a)\boldsymbol{\nu}^a(x|\boldsymbol{\theta}^a)^T\boldsymbol{w}^a\mathrm{d}x = (\boldsymbol{w}^a)^T H^{a,a}(\boldsymbol{\theta}^a)\boldsymbol{w}^a, \tag{3.28}$$

$$J^{a,b}(\Theta^a, \Theta^b) = \int (\boldsymbol{w}^a)^T\boldsymbol{\nu}^a(x|\boldsymbol{\theta}^a)\boldsymbol{\nu}^b(x|\boldsymbol{\theta}^b)^T\boldsymbol{w}^b\mathrm{d}x = (\boldsymbol{w}^a)^T H^{a,b}(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b)\boldsymbol{w}^b, \tag{3.29}$$

$$J^{b,b}(\Theta^b) = \int (\boldsymbol{w}^b)^T\boldsymbol{\nu}^b(x|\boldsymbol{\theta}^b)\boldsymbol{\nu}^b(x|\boldsymbol{\theta}^b)^T\boldsymbol{w}^b\mathrm{d}x = (\boldsymbol{w}^b)^T H^{b,b}(\boldsymbol{\theta}^b)\boldsymbol{w}^b, \tag{3.30}$$

where:

$$\begin{aligned} [H^{a,a}]_{i,j} &= \nu(\mu_i^a|\mu_j^a, \Sigma_i^a + \Sigma_j^a), \\ [H^{b,b}]_{k,l} &= \nu(\mu_k^b|\mu_l^b, \Sigma_k^b + \Sigma_l^b), \\ [H^{a,b}]_{i,k} &= \nu(\mu_i^a|\mu_k^b, \Sigma_i^a + \Sigma_k^b), \end{aligned} \quad \begin{aligned} i, j &= 1, \ldots, n^a, \\ k, l &= 1, \ldots, n^b. \end{aligned} \tag{3.31}$$

Once defined the above quantities, it is possible to obtain a set of D-measures which are functions of such terms. According to [35], a generic LB D-measure for two Gaussian pdfs ν_i and ν_j can always be put in the form:

$$D_{LB}(\nu_i \| \nu_j) = s_{LB}(J^{i,i}, J^{i,j}, J^{j,j}). \tag{3.32}$$

where s_{LB} is a function of the likeness terms. Possessing closed forms in the mixture case is a rare property for a D-measure; as it will be shown, this allows to perform optimization by means, for instance, of gradient descent methods. In this regard, given two Gaussian distributions ν_i and ν, let us define the derivative of a given D_{LB}-measure w.r.t. a generic parameter $\theta = \{\mu, \Sigma^{-1}\}$ of ν:

$$\frac{\partial D_{LB}(\nu_i \| \nu)}{\partial \theta} = c_1^{i,\nu} \frac{\partial J^{i,\nu}}{\partial \theta} + c_2^{i,\nu} \frac{\partial J^{\nu,\nu}}{\partial \theta}, \tag{3.33}$$

where:

$$c_1^{i,\nu} = \frac{\partial s_{LB}(J^{i,i}, J^{i,\nu}, J^{\nu,\nu})}{\partial J^{i,\nu}}, \quad c_2^{i,\nu} = \frac{\partial s_{LB}(J^{i,i}, J^{i,\nu}, J^{\nu,\nu})}{\partial J^{\nu,\nu}}. \tag{3.34}$$

are coefficients specific for each D_{LB}-measure which may depend on all three the likeness terms. The derivative of $J^{i,i}$ w.r.t. θ is, of course, zero. The inverse of the covariance matrix is considered since it can ease the computations when derivatives of Gaussians, or related quantities, are sought.
For the Gaussian case, one obtains:

$$\begin{aligned}
&\frac{\partial J_{i,\nu}}{\partial \mu} = \Sigma^{-1}(\bar{\mu}_{i,\nu} - \mu) J_{i,\nu}, \qquad \frac{\partial J_{\nu,\nu}}{\partial \mu} = 0,\\
&\frac{\partial J_{i,\nu}}{\partial \Sigma^{-1}} = \frac{1}{2}\big(\Sigma - (\bar{\Sigma}_{i,\nu} + (\bar{\mu}_{i,\nu} - \mu)(\bar{\mu}_{i,\nu} - \mu)^T)\big) J_{i,\nu},\\
&\frac{\partial J_{\nu,\nu}}{\partial \Sigma^{-1}} = \frac{1}{2} J_{\nu,\nu} \Sigma.
\end{aligned} \tag{3.35}$$

The partial derivatives above, as it will be discussed in 3.3, result to be useful when solving the so called *barycenter problem.*

Note: the coefficients c_1 and c_2 are independent of the distribution class; the computations here proposed can be employed easily, for instance, for the whole exponential family if one can evaluate the likeness terms.

In the following, a list of D_{LB}-measures are discussed and corresponding closed formulae in the Gaussian case are reported.

Integral Squared Error (Square L2 norm)

Given two distributions p and q, the Integral Squared Error (ISE), also known as Integral Squared Difference (ISD) [37], or Square L_2 norm (L2), is defined as:

$$D_{L2}(p\|q) = \int \big(p(x) - q(x)\big)^2 dx = \|p - q\|_2^2 = \tag{3.36}$$

$$= J^{p,p} - 2J^{p,q} + J^{q,q}. \tag{3.37}$$

Such D-measure satisfies properties (3.1), (3.2) and (3.3). In the ISE case, one gets:

$$c_1^{p,q} = -2, \quad c_2^{p,q} = 1. \tag{3.38}$$

If two Gaussian pdfs ν_i and ν_j are considered, then:

$$D_{L2}(\nu_i\|\nu_j) = \nu(\mu_i|\mu_i, 2\Sigma_i) - 2\nu(\mu_i|\mu_j, \Sigma_i + \Sigma_j) + \nu(\mu_j|\mu_j, 2\Sigma_j), \tag{3.39}$$

and, if a generic $\nu(x|\mu, \Sigma)$ is considered, one obtains:

$$\frac{\partial D_{L2}(\nu_i\|\nu)}{\partial \mu} = \underbrace{-2}_{c_1} \underbrace{\Sigma^{-1}(\bar{\mu}_{i,\nu} - \mu)J^{i,\nu}}_{\frac{\partial J^{i,\nu}}{\partial \mu}} + 0, \tag{3.40}$$

$$\frac{\partial D_{L2}(\nu_i\|\nu)}{\partial \Sigma^{-1}} = \underbrace{-2}_{c_1} \underbrace{\frac{1}{2}\big(\Sigma - (\bar{\Sigma}_{i,\nu} + (\bar{\mu}_{i,\nu} - \mu)(\bar{\mu}_{i,\nu} - \mu)^T)\big)J^{i,\nu}}_{\frac{\partial J^{i,\nu}}{\partial \Sigma^{-1}}} + \underbrace{1}_{c_2} \underbrace{\frac{1}{2}J^{\nu,\nu}\Sigma}_{\frac{\partial J^{\nu,\nu}}{\partial \Sigma^{-1}}}. \tag{3.41}$$

where $\bar{\mu}_{i,\nu}$ and $\bar{\Sigma}_{i,\nu}$ are computed as in (3.26) by considering $\nu_j = \nu$. All the identities used for the computations are also reported in Appendix A.2.
If two Gaussian mixtures p^a and p^b are considered, then:

$$\begin{aligned} D_{L2}(p^a\|p^b) =& J^{a,a} - 2J^{a,b} + J^{b,b} = \\ =& \sum_{i=1}^{n^a}\sum_{j=1}^{n^a} w_i^a w_j^a \nu(\mu_i^a|\mu_j^a, \Sigma_i^a + \Sigma_j^a) + \\ & - 2\sum_{i=1}^{n^a}\sum_{j=1}^{n^b} w_i^a w_j^b \nu(\mu_i^a|\mu_j^b, \Sigma_i^a + \Sigma_j^b) + \\ & + \sum_{i=1}^{n^b}\sum_{j=1}^{n^b} w_i^b w_j^b \nu(\mu_i^b|\mu_j^b, \Sigma_i^b + \Sigma_j^b). \end{aligned} \tag{3.42}$$

By considering the generic parameter $\theta^{b,j} = \{\mu_j^b, (\Sigma_j^b)^{-1}\}$, $j = 1, ..., n^b$, of the j-th Gaussian component ν_j in p^b, and by exploiting (3.40), it is possible to write closed form partial derivatives for the D_{L2} between two mixtures (p^a and p^b) (see Appendix A.2 for a recap of the required identities); nonetheless, the resulting equations are cumbersome and will be omitted, but can be easily obtained by applying the tools here provided. Such a feature is proper of very few D-measures, and the LB family is one of those.

Normalized Integral Squared Error

In the same work [37], Williams proposed even a normalized version of the previous D-measure, namely the Normalized Integral Squared Error (NISE), which, given two distributions p and q, takes the following form:

$$\begin{aligned} D_{NISE}(p\|q) =& \frac{\int (p(x) - q(x))^2 \mathrm{d}x}{\int p(x)^2 \mathrm{d}x + \int q(x)^2 \mathrm{d}x} = \\ =& \frac{J^{p,p} - 2J^{p,q} + J^{q,q}}{J^{p,p} + J^{q,q}} = 1 - \frac{2J^{p,q}}{J^{p,p} + J^{q,q}}. \end{aligned} \tag{3.43}$$

The NISE satisfies properties (3.1), (3.2) and (3.3) and it is confined in the interval $[0, 1]$ (if the pdf support is finite, otherwise $[0, 1)$). In the NISE case, one gets:

$$c_1^{p,q} = \frac{-2}{J^{p,p} + J^{q,q}}, \quad c_2^{p,q} = \frac{2J^{p,q}}{(J^{p,p} + J^{q,q})^2}. \tag{3.44}$$

If two Gaussian pdfs ν_i and ν_j are considered, one gets:

$$D_{NISE}(\nu_i\|\nu_j) = 1 - \frac{\nu(\mu_i|\mu_j, \Sigma_i + \Sigma_j)}{\nu(\mu_i|\mu_i, 2\Sigma_i) + \nu(\mu_j|\mu_j, 2\Sigma_j)}. \tag{3.45}$$

As done for the D_{L2}, one can easily write the NISE between two mixtures p^a and p^b by substituting the single Gaussian likeness terms with the ones computed for GMs, as in 3.28. Moreover, by exploiting (3.35), one can write partial derivatives to obtain gradient for both cases of single Gaussian pdfs and GMs.

Total Square Loss

Another D-measure similar to the ISE and the NISE is the Total Squared Loss (TSL) which has been considered in the literature for shape retrieval applications [38]. Given two distributions p and q over $\mathbb{R}^d$, the TSL is defined as follows:

$$D_{TSL}(p\|q) = \frac{\int (p(x) - q(x))^2 \mathrm{d}x}{\sqrt{1 + 4\int q(x)^2 \mathrm{d}x}} = \frac{J^{p,p} - 2J^{p,q} + J^{q,q}}{\sqrt{1 + 4J^{q,q}}}. \tag{3.46}$$

This D-measure satisfies properties (3.1) and (3.2). Moreover, for the TSL we get:

$$c_1^{p,q} = \frac{-2(1 + 4J^{q,q})}{1 + 4J^{q,q}\sqrt{1 + 4J^{q,q}}}, \quad c_2^{p,q} = \frac{1 - 2(J^{p,p} - 2J^{p,q} - J^{q,q})}{1 + 4J^{q,q}\sqrt{1 + 4J^{q,q}}}. \tag{3.47}$$

As done for the ISE and the NISE, one can write closed forms for both single Gaussian pdfs and GMs, by substituting the likeness terms as discussed.

Cauchy-Schwarz Divergence

One really interesting D-measure, which will be discussed further in this work, is the Cauchy-Schwarz Divergence (CSD) [35, 39, 40]. Given two distributions p and q over $\mathbb{R}^d$, the CSD is defined as follows:

$$D_{CS}(p\|q) = -\log \frac{\int p(x)q(x)\mathrm{d}x}{\sqrt{\int p(x)^2 \mathrm{d}x \int q(x)^2 \mathrm{d}x}}. \tag{3.48}$$

This D-measure satisfies properties (3.1), (3.2) and (3.3); regarding its partial derivatives, one gets:

$$c_1^{p,q} = -\frac{1}{J^{p,q}}, \quad c_2^{p,q} = \frac{1}{2J^{q,q}}. \tag{3.49}$$

In the literature, such D-measure has been used indirectly in [41], where the Correlation Measure (CM) (in the formulation above corresponding to the argument of the logarithm) has been used as *similarity measure* rather than dissimilarity. As it will be discussed, the CSD bears strong similarities with the Bhattacharyya Distance (BD), defined in the next section.
Given two Gaussian pdfs ν_i and ν_j, the CSD takes the following form:

$$D_{CS}(\nu_i\|\nu_j) = \frac{1}{2}\log\frac{|\Sigma_i + \Sigma_j|}{|\Sigma_i|^{\frac{1}{2}}|\Sigma_j|^{\frac{1}{2}}} - \frac{d}{2}\log 2 + \frac{1}{2}(\mu_i - \mu_j)^T(\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j). \tag{3.50}$$

By substituting the likeness terms for the mixture case, one obtains easily the CSD between two GMs.

Jensen-Renyi Quadratic Divergence

One more measure in the LB family is the Jensen-Renyi Quadratic Divergence (JR2D), which is a special case of the *Jensen-Renyi divergences* [42] admitting a closed form for mixtures.

Given two distributions p and q over $\mathbb{R}^d$, the *Jensen-Renyi-*α divergence is defined as follows:

$$D^{\alpha}_{JRD}(p\|q) = H_{\alpha}\left(\frac{p+q}{2}\right) - \frac{H_{\alpha}(p) + H_{\alpha}(q)}{2}, \tag{3.51}$$

with $\alpha \in (-\infty, +\infty)$ and where:

$$H_{\alpha}[p] = \frac{1}{1-\alpha}\log\int p(x)^{\alpha}\mathrm{d}x, \tag{3.52}$$

is the α-Renyi differential entropy [43], which results so be concave in the interval $\alpha \in (0, 1)$, and neither concave or convex in $\alpha \in (-\infty, 0) \cup (1, +\infty)$. The Renyi entropy is a generalization of the Shannon entropy (2.92), and it holds that:

$$\lim_{\alpha\to 1} H_{\alpha}[p] = -\int p(x)\log p(x)\mathrm{d}x. \tag{3.53}$$

In [44], the case $\alpha = 2$ (the Renyi entropy for $\alpha = 2$ takes the name of *collision entropy*) of (3.51) has been considered for a point-set registration problem; in this regard, the JR2D takes the following form:

$$D^{2}_{JRD}(p\|q) = -\log\frac{J^{p,p} + 2J^{p,q} + J^{q,q}}{4\sqrt{J^{p,p}J^{q,q}}}. \tag{3.54}$$

As for all the previous D-measures, it is possible to provide closed forms for the Gaussian pdf and GM case; nonetheless, the JR2D is an ill conditioned measure, since it can take negative values if the considered Gaussians are nearly degenerate in the covariance. For completeness, though, here are reported the coefficients necessary to evaluate the corresponding partial derivatives:

$$c_1^{p,q} = -\frac{4}{2J^{q,q}\left(J^{p,p} + 2J^{p,q} + J^{q,q}\right)}, \quad c_2^{p,q} = \frac{J^{p,p} + 2J^{p,q} - J^{q,q}}{2J^{q,q}\left(J^{p,p} + 2J^{p,q} + J^{q,q}\right)}. \tag{3.55}$$

The optimization of this D-measure, as already mentioned, yields often degenerate solutions associated to negative values of the dissimilarity itself; for this reason, it will not be taken into account anymore in the remainder of this work.

3.1.5 Chernoff α-Divergences

In this section are reported two classes of divergences which fall under the name of Chernoff α-divergences.

Chernoff α-Coefficient

Given two distributions p and q over $\mathbb{R}^d$, the Chernoff α-coefficient (CC) is defined as follows:

$$c_\alpha(p,q) = \int p^\alpha(x) q^{1-\alpha}(x)\mathrm{d}x, \tag{3.56}$$

where $\alpha \in (-\infty, +\infty)$, and $c_\alpha(p,q) \in [0,1] \quad \forall\alpha$. This coefficient is a *similarity* measure between two distributions and, more in detail, it corresponds to the *geometric mean* between p and q.
For $\alpha = \frac{1}{2}$, one obtains the so called Bhattacharyya Coefficient (BC), defined as follows:

$$c_B(p,q) = \int \sqrt{p(x)q(x)}\mathrm{d}x. \tag{3.57}$$

Let us consider the Gaussian case; then the following identities hold:

$$\begin{aligned} \nu^\alpha &= \frac{(2\pi)^{\frac{d}{2}(1-\alpha)}|\Sigma|^{\frac{1-\alpha}{2}}}{\alpha^{\frac{d}{2}}}\nu\left(x|\mu, \tfrac{1}{\alpha}\Sigma\right), \\ \nu^{1-\alpha} &= \frac{(2\pi)^{\frac{d}{2}\alpha}|\Sigma|^{\frac{\alpha}{2}}}{(1-\alpha)^{\frac{d}{2}}}\nu\left(x|\mu, \tfrac{1}{1-\alpha}\Sigma\right). \end{aligned} \tag{3.58}$$

Given two Gaussian distributions $\nu_i = \nu(x|\mu_i, \Sigma_i)$ and $\nu_j = \nu(x|\mu_j, \Sigma_j)$, the Chernoff coefficient takes the following form:

$$c_\alpha(\nu_i,\nu_j) = \left(\frac{|\bar{\Sigma}^\alpha_{i,j}|}{|\Sigma_i|^\alpha|\Sigma_j|^{1-\alpha}}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mu_i-\mu_j)^T(\widetilde{\Sigma}^\alpha_{i,j})^{-1}(\mu_i-\mu_j)\right), \tag{3.59}$$

where

$$\begin{aligned} \bar{\Sigma}^\alpha_{i,j} &= \left(\alpha\Sigma_i^{-1} + (1-\alpha)\Sigma_j^{-1}\right)^{-1}, \\ \widetilde{\Sigma}^\alpha_{i,j} &= \tfrac{1}{\alpha}\Sigma_i + \tfrac{1}{1-\alpha}\Sigma_j, \\ \bar{\mu}^\alpha_{i,j} &= \bar{\Sigma}^\alpha_{i,j}\left(\alpha\Sigma_i^{-1}\mu_i + (1-\alpha)\Sigma_j^{-1}\mu_j\right), \end{aligned} \tag{3.60}$$

and

$$\begin{aligned}\nu_i^{\alpha}(x)\cdot\nu_j^{1-\alpha}(x) =& c_{\alpha}(\nu_i,\nu_j)\cdot\nu(x|\bar{\mu}_{i,j}^{\alpha},\bar{\Sigma}_{i,j}^{\alpha}),\\ \int\nu_i^{\alpha}(x)\cdot\nu_j^{1-\alpha}(x)\mathrm{d}x =& c_{\alpha}(\nu_i,\nu_j)\underbrace{\int\nu(x|\bar{\mu}_{i,j}^{\alpha},\bar{\Sigma}_{i,j}^{\alpha})\mathrm{d}x}_{=1}.\end{aligned} \tag{3.61}$$

For $\alpha=\frac{1}{2}$, the Bhattacharyya coefficient takes the following form:

$$c_B(\nu_i,\nu_j)=\frac{|\Sigma_i|^{\frac{1}{4}}|\Sigma_j|^{\frac{1}{4}}}{|\frac{\Sigma_i+\Sigma_j}{2}|^{\frac{1}{2}}}\exp\Bigg(-\frac{1}{8}(\mu_i-\mu_j)^T\Bigg(\frac{\Sigma_i+\Sigma_j}{2}\Bigg)^{-1}(\mu_i-\mu_j)\Bigg). \tag{3.62}$$

When mixtures are considered, such coefficient does not admit a closed form. **Note:** in the Gaussian case, the Chernoff coefficient bears some similarities with the cross-likeness term (eq. (3.24)); the former is derived from the geometric mean of the parameters of two Gaussian densities, while the latter can be seen as the *resistor parallel* equivalent.
Analogous derivatives as the ones in (3.35) can be obtained for the Chernoff α-coefficient, but are omitted due to their cumbersome forms.

By exploiting the Chernoff coefficient, it is possible to define two kinds of divergences, which are reported in the following.

I° kind divergences

Given two distributions p and q over $\mathbb{R}^d$, the Chernoff α-Divergence of the I° kind (CD$'_{\alpha}$) is defined as:

$$D'_{\alpha}(p\|q)=\frac{1}{\alpha(1-\alpha)}(1-c_{\alpha}(p,q)). \tag{3.63}$$

An interesting fact regarding this class of dissimilarities is that, for several values of α, some known divergences are obtained:

$$\begin{aligned}
D'_\alpha(p\|q)\big|_{\alpha=-1} = D_P(p\|q) =& \frac{1}{2}\bigg(\int \frac{q(x)^2}{p(x)}\mathrm{d}x - 1\bigg), && \text{Pearson } \chi^2 \\
\lim_{\alpha\to 0} D'_\alpha(p\|q) = D_{RKL}(p\|q) =& \int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}x, && \text{Reverse KLD} \\
D'_\alpha(p\|q)\big|_{\alpha=0.5} = 4D_{H2}(p\|q) =& 4\big(1 - \int \sqrt{p(x)q(x)}\mathrm{d}x\big), && \text{Square Hellinger} \\
\lim_{\alpha\to 1} D'_\alpha(p\|q) = D_{FKL}(p\|q) =& \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x, && \text{Forward KLD} \\
D'_\alpha(p\|q)\big|_{\alpha=2} = D_N(p\|q) =& \frac{1}{2}\bigg(\int \frac{p(x)^2}{q(x)}\mathrm{d}x - 1\bigg). && \text{Neyman } \chi^2
\end{aligned} \tag{3.64}$$

When the Gaussian case is considered, it is convenient to restrict the α interval to $[0, 1]$, since the Pearson χ^2 Divergence (PD) and the Neyman χ^2 Divergence (ND) are well defined only for a particular subset of $\mathcal{N}^d$, which we recall to be the space of all the d-dimensional Gaussian densities; in this regard, the closed forms are provided only for the Square Hellinger Distance (H2) (the FKLD and RKLD have been already defined in 3.1.1). Given two Gaussian densities ν_i and ν_j, the Square Hellinger Distance is defined as follows:

$$D_{H2}(\nu_i\|\nu_j) = 1 - c_B(\nu_i, \nu_j), \tag{3.65}$$

where $c_B(\nu_i, \nu_j)$ has been defined in (3.62). The H2 satisfies the properties (3.1), (3.2) and (3.3).

When mixtures are considered, none of the I° kind divergences admit a closed form.

II° kind divergences

Given two distributions p and q over $\mathbb{R}^d$, the Chernoff α-Divergence of the *II°* kind (CD''_α) is defined as:

$$D''_\alpha(p\|q) = -\log c_\alpha(p,q). \tag{3.66}$$

For $\alpha = \frac{1}{2}$ one obtains the Bhattacharyya distance. If two Gaussian densities ν_i and ν_j are considered, the Bhattacharyya distance has the following closed form:

$$D_B(\nu_i\|\nu_j) = \frac{1}{2}\log\frac{|\Sigma_i+\Sigma_j|}{|\Sigma_i|^{\frac{1}{2}}|\Sigma_j|^{\frac{1}{2}}} - \frac{d}{2}\log 2 + \frac{1}{4}(\mu_i-\mu_j)^T(\Sigma_i+\Sigma_j)^{-1}(\mu_i-\mu_j). \tag{3.67}$$

This D-measure satisfies properties (3.1), (3.2) and (3.3). One can notice the similarity between the Bhattacharyya and the Cauchy-Schwarz (defined in (3.48)): the quadratic form in the CSD is multiplied by a factor of two if compared to the BD.

When mixtures are considered, none of the divergences in this class possesses a closed form.

3.1.6 Other Dissimilarity Measures

All the D-measures above are just a fraction of the ones present in the literature like, for instance, the *Renyi Divergence* [43] or the *Jensen-Shannon Divergence* [45]. For the goals of this work, though, the author wanted to list a large pool, with corresponding closed formulas in the Gaussian case, in order to provide several off-the-shelf alternatives for the framework presented in Chap. 4. To know more about statistical distances, a reference name is Frank Nielsen, who has given many contributions to the so called Information Geometry field, and who has addressed and formalized many related topics.

3.2 Mixture Reduction: Problem Formulation

When the uncertainty of a process is described by a mixture of densities, it can happen that the corresponding number of components grows significantly; in those cases, managing such a representation may be computationally intractable and approximations should be introduced. Formally, given a mixture p^a with n^a components, the mixture reduction problem consists in finding a mixture p^b, with $n^b < n^a$ components, which minimizes a desired dissimilarity D from p^a. For a chosen D-measure, the theoretical solution to this problem is given by:

$$\Theta^{b^*} = \underset{\Theta^b \in \mathcal{H}_{n^b}}{\arg\min}\, D(p(x|\Theta^a) \| p(x|\Theta^b)). \tag{3.68}$$

This is in general a complex, non-convex constrained nonlinear optimization problem, which does not admit a closed form solution. Moreover, when dealing with mixtures, only very few D-measures admit a closed form, hence making the problem defined as above often analytically intractable. In this regard, all the existing approaches in the literature are based on heuristics which try to tackle the MRP by minimizing different D-measures in the same reduction pipeline. Procedures like that lead in general to *inconsistency* issues, which basically consist in preferring ease of computation rather than seeking better solutions w.r.t. a desired D-measure.

3.2.1 Consistency in Mixture Reduction

A mixture reduction procedure is said to be *consistent*[1] if all actions and steps involved in the process are done according to a single D-measure. As discussed in [46], consistency is not an obvious feature of reduction algorithms, and in fact most algorithms proposed in the literature are not consistent. The problem as defined as in (3.68) requires a dissimilarity measure between mixtures to be minimized; from an optimization point of view, then, it would be correct to try to perform any kind of minimization according to such D-measure. Hence, given the problem (3.68), performing consistent actions is expected to yield superior solutions w.r.t. the given dissimilarity if compared to the inconsistent case; for this reason, one should try, when possible, to perform all the greedy reduction/refinement steps accordingly. Nonetheless, given the lack of closed forms when addressing the MRP, it might be necessary to often resort either to analytically tractable approximations of the involved quantities or to numerical alternatives when particular quantities are sought. Some numerical tests to discuss further the concept of consistency will be proposed in Chapter 5.

[1] Interchangeably the terms *congruent* or *coherent* will be used.

3.2.2 A Common Approach

As mentioned, the solution of (3.68) is in general analytically intractable, hence numerical solutions have to be sought. Due to the presence of many local minima, the outcome of the numerical iterative solution of (3.68) strongly depends on the choice of the starting point. For this reason, most MR algorithms proposed in the literature include a preliminary iterative reduction phase of the original mixture, where at each step the corresponding size is reduced by one or more components, according to some *greedy* criterion, until the required size n^b is reached. The reduced mixture obtained in this *Greedy Reduction* phase can possibly serve as a starting point for a *Refinement* phase, aimed at minimizing further the dissimilarity while keeping constant the size of the reduced mixture.

3.2.3 Greedy Reduction

The basic actions in the Greedy Reduction are *pruning*, i.e. removal from the mixture of one or more components, or *merging* of two or more components into one component.

Many greedy reduction algorithms in the literature employ pruning for reducing the complexity of mixture representations (e.g. [13, 37] and many more); nonetheless, pruning is a *destructive* practice, in the sense that the information removed by means of it is completely lost. Thus said, in the literature have been proposed several criteria to perform pruning of mixture components, for instance:

- pruning by weight importance: select one or more components associated to the smallest weights and remove those from the mixture; a weight renormalization is then performed for the remaining components.

- pruning by least dissimilarity introduced: some cost-function based pruning criteria (e.g. [37]) evaluate the cost of removing a component from the mixture in terms of dissimilarity introduced. If such a cost is acceptable, the component is pruned and a weight renormalization follows. Note that, as it will be further discussed, pruning is something strictly linked with the peculiarities of a D-measure; there are many

D-measures more *exclusive*[2] than others, hence more prone to pruning-like behaviors.

- pruning by statistical importance: consider the statistics of the weights (e.g., the cdf) and prune components associated to a low percentile (e.g. [47]).

Alternatively to pruning, one can perform merging; in this work, two kinds of merging actions will be considered, that is to merge mixture components by means of *barycenter* or by means of Best Single Density Approximation (BSDA). In order to deal intuitively with those two quantities, it is necessary to recall the concepts of *normalized* and *unnormalized* sub-mixtures, defined in Sec. 2.8.1.

Any MR algorithm that performs merging of components of a given mixture $p(x|\Theta)$ acts as follows:

1) a sub-mixture $p(x|\Theta_{\mathcal{I}}) \subset p(x|\Theta)$ is selected according to some criterion;

2) the selected sub-mixture $p(x|\Theta_{\mathcal{I}})$ is replaced in $p(x|\Theta)$ by a single density, chosen according to some criterion, with weight $\boldsymbol{w}_{\mathcal{I}}^T \mathbb{1}_{\bar{n}}$ (so that the sum of the weights of the reduced mixture remains one).

The criteria adopted in the two steps above should be consistent with the same D-measure in order to obtain a reduced mixture not too dissimilar from the original one, according to the chosen D-measure. A consistent pipeline of reduction actions in a MR algorithm is achieved when the criteria in the pruning and merging steps are based on the same D-measure used in the refinement phase.

3.2.4 Best Single Density Approximation

Let us consider a mixture $p = \boldsymbol{w}^T \boldsymbol{q} \in \mathcal{Q}_{\text{mix}}$ of size n which elements are in a class $\mathcal{Q}$; the Best Single Density Approximation (BSDA) $q^* \in \mathcal{Q}'$ of p, according to a given D-measure, is defined as:

$$q^* = \arg\min_{q \in \mathcal{Q}'} D(\boldsymbol{w}^T \boldsymbol{q} \| q), \tag{3.69}$$

[2] With the term exclusive are denoted D-measures which tend to neglect low-density regions of a distributions in favor of preserving other features, e.g., the main peaks of a multimodal density.

that is the density which is the least dissimilar from the mixture p according to the dissimilarity D. In general, it is of interest to find the BSDA distribution in the same class, that is $\mathcal{Q}' = \mathcal{Q}$. The BSDA, as the name says, is a density, hence, for its computation, it is required to work either with mixtures as a whole or normalized sub-mixtures. If the BSDA is selected as merging method in a mixture reduction algorithm as previously discussed, one can evaluate such quantity correctly by at first normalizing the selected sub-mixture and, after computing the associated BSDA, by assigning to it the weight obtained as the sum of the involved components weights. For most of the D-measures presented in Sec. 3.1, the BSDA computation can not be done in a closed form, mostly because the dissimilarity between a mixture and a single density is usually analytically intractable. For the goals of this work, the BSDA will not represent the main approach for merging components; instead, the concept of *barycenter* of a set of weighted densities will result to be a core component in most of the reported algorithms. In [35], a discussion regarding the BSDA computation for the LB family in the Gaussian case is reported.

3.3 The Barycenter Problem

Given a vector $\boldsymbol{w} \in \mathbb{R}^n_+$ and a set $\boldsymbol{q}$ of n distributions q_i, $i \in [1:n]$, in the class (family) $\mathcal{Q}$ of pdfs, the Average Dissimilarity Function (ADF) of the weighted set $(\boldsymbol{w}, \boldsymbol{q})$ (also denoted $\{w_i q_i\}_{i=1}^n$), from a given (generic) pdf q, according to a given D-measure, is defined as

$$m_D(q \,|\, \boldsymbol{w}, \boldsymbol{q}) = \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i D(q_i \| q). \tag{3.70}$$

If $\boldsymbol{w} \in \Delta^{n-1}$, then $\boldsymbol{w}^T \mathbb{1}_n = 1$, and the weighted set $\{w_i q_i\}_{i=1}^n$ can represent a mixture, which is a pdf, defined as $\boldsymbol{w}^T \boldsymbol{q} \in \mathcal{Q}_{\text{mix}}$ (a short notation for $\sum_{i=1}^n w_i q_i$).

The barycenter $\hat{q}$ of the weighted set $(\boldsymbol{w}, \boldsymbol{q}) = \{w_i q_i\}_{i=1}^n$ is defined as the distribution q that minimizes the ADF (3.70). Mostly, it is of interest to find the barycenter in the same class $\mathcal{Q}$ of distributions q_i in the set $\boldsymbol{q}$. Thus, the barycenter, denoted $\hat{q}$, is defined as

$$\hat{q} = \arg\min_{q \in \mathcal{Q}} m_D(q|\boldsymbol{w}, \boldsymbol{q}) \triangleq \bar{\Phi}_D(\boldsymbol{w}, \boldsymbol{q}). \tag{3.71}$$

For a matter of notation, the function $\bar{\Phi}_D(\cdot)$[3] has been defined as the operator returning the D-barycenter(s) of either a set of weighted densities or a mixture/intensity. It is straightforward to see that the denominator $\boldsymbol{w}^T \mathbb{1}_n$ in (3.70) does not play any role in the computation of the barycenter.

As a general rule for the Gaussian case, to compute the barycenter $\hat{\nu} = \nu(x|\hat{\mu}, \widehat{\Sigma})$ of a set of weighted Gaussian components $(\boldsymbol{w}, \boldsymbol{\nu})$ one can proceed by evaluating the partial derivatives of (3.70) w.r.t. the parameters $\theta = \{\mu, \Sigma^{-1}\}$ of a generic component $\nu(x|\mu, \Sigma)$ as:

$$\begin{aligned}
\frac{\partial m_D(\nu|\boldsymbol{w}, \boldsymbol{\nu})}{\partial \mu} &= \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \frac{\partial D\big(\nu_i \| \nu(\cdot|\mu, \Sigma)\big)}{\partial \mu} = 0, \\
\frac{\partial m_D(\nu|\boldsymbol{w}, \boldsymbol{\nu})}{\partial \Sigma^{-1}} &= \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \frac{\partial D\big(\nu_i \| \nu(\cdot|\mu, \Sigma)\big)}{\partial \Sigma^{-1}} = 0.
\end{aligned} \tag{3.72}$$

Nonetheless, as it will be discussed, only for very few D-measures such system of partial derivatives admits a closed form solution; when this does not happen, one can resort either to gradient descent optimization or to Fixed-Point Iteration (FPI) algorithms. As discussed in [48], the latter approach overcomes the former when the barycenter problem is addressed, given that gradient descent has to be constrained to preserve the admissible domain of the parameters. For instance, the covariance matrix of a Gaussian density has to belong to S^d_{++}, which is often not an easy constraint to satisfy. Moreover, one has to find a suitable gradient step size, which can be a rather critical choice when computing barycenters; for a more detailed discussion, see [48]. For the remainder of this work, when (3.72) does not admit a closed form solution, FPI algorithms will be considered.

Definition 3.3.1. (Associativity of barycenters) *For a given D-measure and given class $\mathcal{Q}$ of distributions, the barycenters are said to be* associative *if for any given weighted set $\{w_i q_i\}_{i=1}^n$ of distributions $q_i \in \mathcal{Q}$, and for any given pair of disjoint subsets $\mathcal{I}_1$ and $\mathcal{I}_2$ of the interval* $[1\!:\!n]$*, the following identity*

[3] Such a function can take either sets of weighted components, written either in the form $(\boldsymbol{w}, \boldsymbol{q})$ or $\{w_i q_i\}_{i=1}^n$ (e.g. $\bar{\Phi}_D(\boldsymbol{w}, \boldsymbol{q})$, $\bar{\Phi}_D(\{w_i q_i\}_{i=1}^n)$), or mixture/intensities, either in normalized or unnormalized forms (e.g. $\bar{\Phi}_D(\boldsymbol{w}^T \boldsymbol{q})$, $\bar{\Phi}_D(\tilde{w}_i q_i + \tilde{w}_j q_j)$).

holds true:

$$\begin{aligned}&\bar{\Phi}_D\big(\{w_i q_i\}_{i\in\mathcal{I}_1\cup\mathcal{I}_2}\big) = \bar{\Phi}_D\big(\{\bar{w}_{\mathcal{I}_1}\hat{q}_{\mathcal{I}_1}, \bar{w}_{\mathcal{I}_2}\hat{q}_{\mathcal{I}_2}\}\big),\\ &\textit{where} \quad \hat{q}_{\mathcal{I}_j} = \bar{\Phi}_D\big(\{w_i q_i\}_{i\in\mathcal{I}_j}\big), \quad \bar{w}_{\mathcal{I}_j} = \sum_{i\in\mathcal{I}_j} w_i, \quad j \in \{1,2\}.\end{aligned} \tag{3.73}$$

An equivalent form for the associativity property (3.73), that uses unnormalized mixtures as argument of the function $\bar{\Phi}_D(\cdot)$, instead of weighted sets, is the following:

$$\hat{q}_{\mathcal{I}_1\cup\mathcal{I}_2} = \bar{\Phi}_D\Big(\sum_{i\in\mathcal{I}_1\cup\mathcal{I}_2} w_i q_i\Big) = \bar{\Phi}_D\big(\bar{w}_{\mathcal{I}_1}\hat{q}_{\mathcal{I}_1} + \bar{w}_{\mathcal{I}_2}\hat{q}_{\mathcal{I}_2}\big). \tag{3.74}$$

Following is reported another fundamental property for the D-barycenter of a set of weighted densities.

Definition 3.3.2. *A family of distributions $\mathcal{Q}$ and a D-measure are said to satisfy the Average Barycentral Triangular Identity (ABTI) property if for any given weighted set $\{w_i q_i\}_{i=1}^n$ the following holds true*

$$\sum_{i=1}^n w_i D(q_i\|q) = \sum_{i=1}^n w_i D(q_i\|\hat{q}) + \Big(\sum_{i=1}^n w_i\Big) D(\hat{q}\|q), \quad \forall q \in \mathcal{Q}, \tag{3.75}$$

where $\hat{q} = \bar{\Phi}_D(\{w_i q_i\}_{i=1}^n)$.

The identity (3.75) can be rewritten in the form

$$\sum_{i=1}^n w_i D(q_i\|q) = \sum_{i=1}^n w_i\big(D(q_i\|\hat{q}) + D(\hat{q}\|q)\big), \quad \forall q \in \mathcal{Q}, \tag{3.76}$$

that justifies the name ABTI. Indeed, in general, given a triple $(q_i, \hat{q}, q)$ (a triangle in $\mathcal{Q}$), where q_i is any component of a weighted set $\{w_i q_i\}_{i=1}^n$, $\hat{q}$ is its barycenter, and q any distribution in $\mathcal{Q}$, we have that $D(q_i\|q)$ can be larger, equal, or smaller than the sum $D(q_i\|\hat{q}) + D(\hat{q}\|q)$, i.e.

$$D(q_i\|q) \lesseqqgtr D(q_i\|\hat{q}) + D(\hat{q}\|q). \tag{3.77}$$

The family $\mathcal{Q}$ and the D-measure satisfy the ABTI property if, for any given weighted set $\{w_i q_i\}_{i=1}^n$, averaging both the left and right hand sides of (3.77), using the weight set $\boldsymbol{w} = \{w_i\}_{i=1}^n$, the identity (3.76) is obtained.

Remark 1. *D-measures that satisfy the triangular inequality*

$$D(q_1\|q_3) \leq D(q_1\|q_2) + D(q_2\|q_3), \quad \forall q_1, q_2, q_3 \in \mathcal{Q}, \tag{3.78}$$

are such that

$$\sum_{i=1}^{n} w_i D(q_i\|q) \leq \sum_{i=1}^{n} w_i \big(D(q_i\|\hat{q}) + D(\hat{q}\|q)\big), \quad \forall q \in \mathcal{Q}, \tag{3.79}$$

and therefore they do not satisfy the ABTI property (3.76)*; none of the D-measures listed in this work satisfy the triangular inequality property.*

3.3.1 Discussion about the features of a D-measure

As mentioned, there exists a broad range of D-measures, and they all exhibit different features[4], both in terms of analytical properties and peculiarities. In this regard, it is often not obvious which dissimilarity would be more suitable for a problem of interest, since, as it will be discussed, the corresponding choice should always take into account the problem structure and related desired outcomes. A discussion regarding the features of D-measures is not the main goal of this work, since it is not an easy task and there is not much literature available. In any case, a preliminary discussion regarding this topic is required to better understand some considerations which will follow in the remainder of this dissertation. Together with some observations derived from the experience of the author, terms like *inclusive* or *exclusive* will be often used when referring to the features of a given D-measure. Such terms have been coined in the work of Minka [49] to describe the peculiarities of a D-measure; to the best of the author's knowledge, not much of literature has been produced regarding such a topic, probably due to the fact that it is not easy to identify particular metrics to evaluate such a thing. Nonetheless, after a broad campaign of tests, it has been noticed that investigating the outcome of a barycenter (or BSDA) computation can provide some interesting insights about a given D-measure. In this regard, following will be reported a series of "grid" tests, that is the function (3.70) will be evaluated over a grid to be visualized as function of parameters. Moreover, potential

[4] With the word feature, the author is denoting the peculiarities of a dissimilarity measure, since, as it will be discussed more in detail in this dissertation, two D-measures can exhibit very different behaviors when applied to the same problem.

closed form solutions, or FPI algorithms, will be reported for a broad range of D-measures, by discussing the corresponding properties. Since in this work the barycenter will be favored over the BSDA as merging method, the analysis will be restricted to (3.70); nonetheless, some insights regarding the BSDA will be provided for each D-measure.

Let us consider the set of 1-dimensional Gaussian densities parameterized as follows:

$$\begin{aligned} d = 1, \qquad \boldsymbol{w} &= \{0.2, 0.3, 0.5\}, \\ \boldsymbol{\mu} = \{-1, 0, 3.5\}, \qquad \boldsymbol{\Sigma} &= \{0.05, 0.1, 0.15\}. \end{aligned} \tag{3.80}$$

3.3.2 Forward Kullback-Leibler Divergence Barycenter

Only for the D_{FKL} and the D_{RKL} the barycenter computation will be addressed for the gammas and inverse-Wisharts in addition to the Gaussian case. Given (3.80), the corresponding D_{FKL}-bar surface is reported in Fig.3.1. The D_{FKL} admits a closed form for the barycenter computation which is

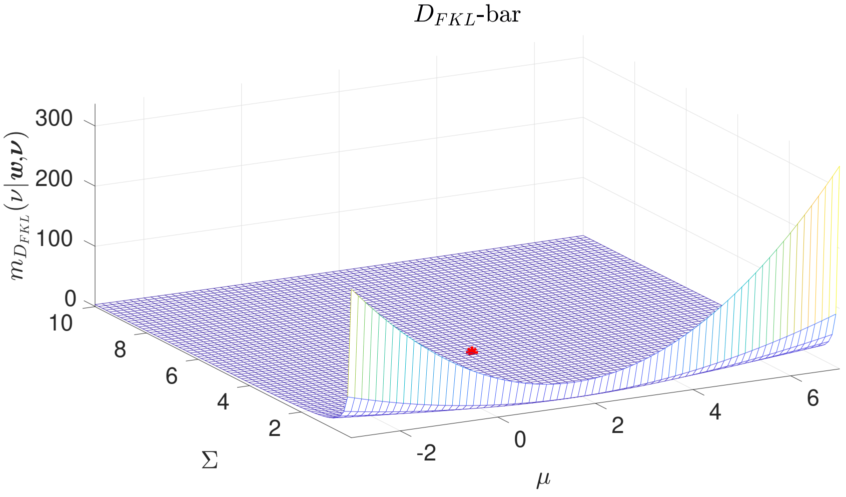

Figure 3.1: D_{FKL}-bar: $\hat{\mu} = 1.5500$, $\widehat{\Sigma} = 4.0375$.

known in the literature as *moment-preserving merge* [50], or *moment match-*

ing, which formulae are the following:

$$\hat{\mu} = \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \mu_i \triangleq \mu_{MP}, \tag{3.81a}$$

$$\widehat{\Sigma} = \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \big(\Sigma_i + (\mu_i - \hat{\mu})(\mu_i - \hat{\mu})^T\big) \triangleq \Psi_{MP}(\hat{\mu}). \tag{3.81b}$$

The corresponding barycenter will then be $\hat{\nu} = \nu(\cdot|\hat{\mu}, \widehat{\Sigma})$. Moreover, the D_{FKL}-barycenter and the D_{FKL}-BSGA do coincide and the proof is reported in Appendix B.1.1. Such a property results to be really interesting, since it says that the D_{FKL}-barycenter, which has a closed analytical form, is the best approximation one can provide for both a set of weighted Gaussians and the corresponding mixture built as their convex sum. Moreover, the D_{FKL} is a *Bregman divergence*, which, in the barycenter case, admits a unique minimizer [51]. For the Gaussian case, a closed form solution is given, but this might not be true for other distributions in the exponential family; nonetheless, when this happens, the unique solution can be approximated by means of numerical methods with arbitrary accuracy. In this regard, let us consider now a set of weighted inverse-Wishart densities $(\boldsymbol{w}, \boldsymbol{\varphi}) = \{w_i \varphi_i\}_{i=1}^n$; by solving (3.72) in the D_{FKL} case, one obtains the following system:

$$\sum_{m=1}^{d} \psi_0\left(\frac{\hat{v} - d - m}{2}\right) - d \log \frac{\hat{v} - d - 1}{2} = h, \tag{3.82a}$$

$$\widehat{V} = \frac{\hat{v} - d - 1}{2} H^{-1}, \tag{3.82b}$$

where:

$$\begin{aligned} H &= \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \frac{v_i - d - 1}{2} V_i^{-1}, \\ h &= \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \left(\sum_{m=1}^{d} \psi_0\left(\frac{v_i - d - m}{2}\right) - \log |V_i| \right) - \log |H|. \end{aligned} \tag{3.83}$$

Although the system of equations (3.82) is formally a system whose unknown is in a high dimensional space, since $(\widehat{V}, \hat{v}) \in S_{++}^d \times \mathbb{R}_+$, it turns out that it can be solved by simply solving (3.82a), which is a scalar equation in the scalar unknown $\hat{v}$. Once $\hat{v}$ is computed, the matrix unknown $\widehat{V}$ is obtained using (3.82b). In order to find the (unique) solution to (3.82a), one can resort to the *Newton-Rhapson (NR)* algorithm. Let us consider:

$$\begin{aligned} f(v) &= \sum_{m=1}^{d} \psi_0\left(\frac{v-d-m}{2}\right) - d\log\frac{v-d-1}{2} - h, \\ f'(v) &= \frac{1}{2}\sum_{m=1}^{d} \psi_1\left(\frac{v-d-m}{2}\right) - \frac{d}{v-d-1}. \end{aligned} \tag{3.84}$$

Then, the NR iteration is:

$$v^{(k+1)} = v^{(k)} - \frac{f(v^{(k)})}{f'(v^{(k)})}. \tag{3.85}$$

Nonetheless, ψ_0 assumes real values only for positive arguments, hence situations where $v^{(k)} < 2d$ are not admissible in the NR recursion. In this regard, the starting point $v^{(0)}$ is fundamental either for convergence time, or for numerical stability of the root-finding algorithm. In Fig. 3.2 is reported a graphical intuition about the features of $f(v)$. First of all, $f(v)$ is a strictly increasing concave function, hence $f'(v)$ is always positive in the interval $(2d, +\infty)$. If $2d < v < \hat{v}$ is chosen, one gets $f(v) < 0$, and the NR algorithm is expected to provide a uniform convergence towards $\hat{v}$ from below. If instead the recursion is initialized in a value $v > \hat{v}$, $f'(v) \to 0$, while $f(v) > 0$, hence providing significant updates towards values of $v^{(k)}$ which might be smaller than $2d$, which is the case reported in Fig. 3.2. In this regard, in order to have a guaranteed convergence to $\hat{v}$, a suitable initialization could be $v^{(0)} = 2d + \epsilon$, with ϵ small enough (e.g. $\epsilon = 10^{-3}$); such starting point might be far from $\hat{v}$, but it should be noted that the recursion on v is scalar, hence computationally lightweight. Finally, by iterating the NR algorithm until a given desired accuracy is reached, the D_{FKL}-barycenter of the $\mathcal{IW}$ densities is parameterized as $\hat{\varphi} = \varphi(X|\widehat{V}, \hat{v})$, where $\hat{v}$ is the solution provided by the root-finding algorithm, and $\widehat{V}$ is computed accordingly.

As for the Gaussian case, the D_{FKL}-BSDA coincides with the D_{FKL}-barycenter, and this can be proven by doing the same considerations as before.

Let us consider now a set of weighted gamma densities $(\boldsymbol{w}, \boldsymbol{\gamma}) = \{w_i\gamma_i\}_{i=1}^{n}$;

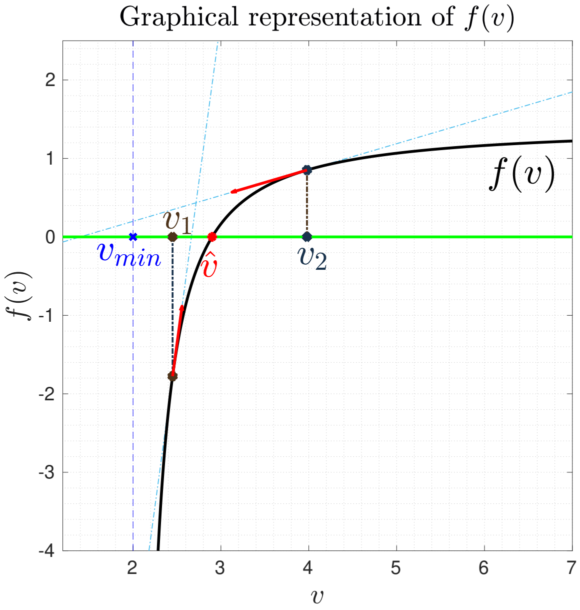

Figure 3.2: Graphical intuition of the $f(v)$ initializations (1D).

by proceeding as before, one obtains the following system:

$$\psi_0(\hat{\kappa}) - \log \hat{\kappa} = g_2, \tag{3.86a}$$

$$\hat{\omega} = \frac{g_1}{\hat{\kappa}}, \tag{3.86b}$$

where:

$$\begin{aligned} g_1 &= \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \kappa_i \omega_i, \\ g_2 &= \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \big(\psi_0(\kappa_i) + \log \omega_i\big) - \log g_1. \end{aligned} \tag{3.87}$$

Even in this case, the system does not admit a closed form solution. Nonetheless, by comparing (3.86) to (3.82), one notices that the system associated to the gamma D_{FKL}-barycenter is a simpler case of the one discussed for the inverse-Wisharts, hence only (3.86a) has to be solved for the scalar $\hat{\kappa}$, and then $\hat{\omega}$ is simply computed using (3.86b); the only main difference is that $\kappa_{min} = 0$. Regarding the initialization, if the digamma series expansion truncated to the second order (see generalized *Puiseux* series) is considered in (3.86), one obtains a second degree polynomial equation from which the accurate initialization $\kappa^{(0)} = \frac{-(1+\sqrt{1-4g_2/3})}{4g_2}$ can be obtained. It can be shown that the discriminant in such formula is always positive. A similar approach can be considered for the inverse-Wishart, but, from numerical tests, it has been observed that for high dimensional problems such a starting point can be rather coarse, and often leads to negative arguments for the digamma function. Once solved the equations (3.86a) and (3.86b), the D_{FKL}-barycenter of the set of weighted gammas is parameterized as $\hat{\gamma} = \gamma(\chi|\hat{\kappa}, \hat{\omega})$. As for the Gaussians and inverse-Wisharts, the D_{FKL}-BSDA and D_{FKL}-barycenter coincide: this property is intrinsic of the D-measure itself, hence can be extended to the whole exponential family.

Are there any similarities between the just discussed formulae and any of the previous sections? By considering again Sec. 2.3, it is rather easy to notice the strong similarity between the MLE of those three densities given a set of samples and the barycenter computation for a set of densities. In fact, an interesting property of the D_{FKL} measure is that, as discussed in 3.5, it is strictly linked to the ML principle. The only differences between the sample data and set of densities cases are the values assumed by the constants and the fact that the densities can have different weights, while the samples all have weight $\frac{1}{n}$. This interesting fact will be exploited later by coupling the EM algorithm together a mixture reduction approach based on the D_{FKL} to find mixture models with a likely number of components.

In general, the D_{FKL} is regarded as an *inclusive* D-measure, in the sense that it tries to preserve the most of the mixture density by spreading the barycenter covariance when the weighted components are spread apart in the space. As it will be discussed, such a property can be problematic in some applications.

D_{FKL}-barycenters in the exponential family and their properties

Let us recall the EF as defined in Sec. 2.2; some properties of D_{FKL}-barycenters of weighted sets of pdfs in an exponential family of distributions $\mathcal{Q}$ with natural parameter $\eta \in \Lambda \subset \mathbb{R}^{n_\eta}$ will be presented in the following. $q(\eta)$ will denotes a pdf in $\mathcal{Q}$, as a shorthand for $q(x|\eta)$. Moreover, when considering a weighted set $(\boldsymbol{w}, \boldsymbol{q})$ of distributions in $\mathcal{Q}$, the short notation q_i is used instead of $q(x|\eta_i)$ to denote a distribution with natural parameter η_i. With respect to the mixture case defined in 2.8, the given parameters are now η (the natural parameters). This change will only affect this section, since it makes more elegant and compact the discussion which will follow.

Proposition 3.3.1. *Let $\mathcal{Q}$ be an exponential family of distributions, with natural parameter $\eta \in \Lambda \subset \mathbb{R}^{n_\eta}$. The D_{FKL}-barycenter of a weighted set $(\boldsymbol{w}, \boldsymbol{q}) = \{w_i q_i\}_{i=1}^n$ of distributions $q_i \in \mathcal{Q}$ is unique, and its natural parameter $\hat{\eta} \in \Lambda$ is such that:*

$$\hat{\eta}: \quad b(\hat{\eta}) = \frac{1}{\bar{w}} \sum_{i=1}^{n} w_i b(\eta_i). \tag{3.88}$$

where $\bar{w} = \boldsymbol{w}^T \mathbb{1}_n$. Since the solution $\hat{\eta}$ of (3.88) *is unique, $\hat{\eta}$ can be written as*

$$\hat{\eta} = b^{-1}\left(\frac{1}{\bar{w}} \sum_{i=1}^{n} w_i b(\eta_i)\right). \tag{3.89}$$

Proof of uniqueness is reported in Appendix B.2.1.

Remark 2. *Taking into account* (3.88)*, one can write:*

$$m_{D_{FKL}}(q(\eta)|\boldsymbol{w}, \boldsymbol{q}) = \sum_{i=1}^{n} w_i \big(\eta_i^T b(\eta_i) - a(\eta_i)\big) - \bar{w}\Big(\eta^T b(\hat{\eta}) - a(\eta)\Big). \tag{3.90}$$

Equation (3.88) implies the **associativity** property for the D_{FKL}-barycenters of weighted sets and mixtures of distributions in an exponential family $\mathcal{Q}$, as proved in the following theorem.

Theorem 3.3.2. *The D_{FKL}-barycenters of weighted sets of pdfs in a class $\mathcal{Q}$ of distributions in the exponential family, are associative.*

Proof of associativity for D_{FKL}-barycenters is reported in Appendix B.2.2.

Another important property possessed by the D_{FKL}-barycenters is the Average Barycentral Triangular Identity (ABTI).

Proposition 3.3.3. *Given $q_1, q_2 \in \mathcal{Q}$, with $\mathcal{Q}$ an exponential family, and two positive weights w_1 and w_2, the following is true*

$$\begin{aligned} &w_1 D_{FKL}(q_1 \| q) + w_2 D_{FKL}(q_2 \| q) = \\ &= w_1 D_{FKL}(q_1 \| \hat{q}_{1,2}) + w_2 D_{FKL}(q_2 \| \hat{q}_{1,2}) + (w_1 + w_2) D_{FKL}(\hat{q}_{1,2} \| q), \quad \forall q \in \mathcal{Q}. \end{aligned} \tag{3.91}$$

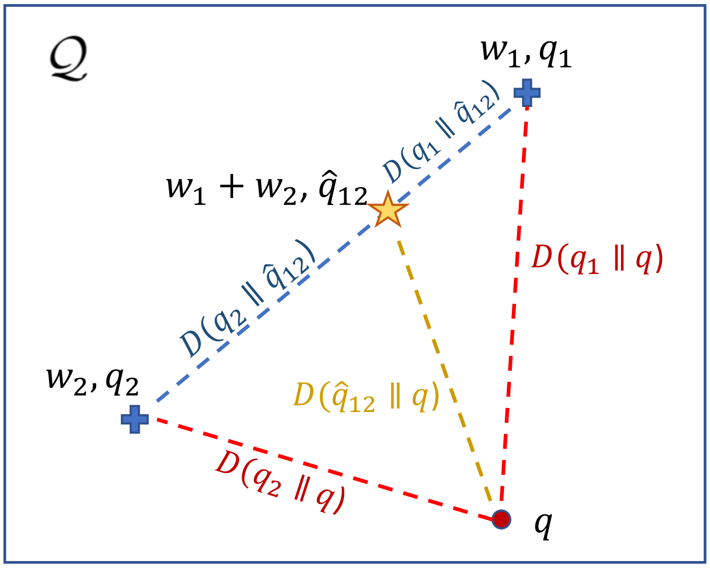
$\mathcal{Q}$
w_1, q_1
$D(q_1 \parallel \hat{q}_{12})$
$w_1 + w_2, \hat{q}_{12}$
$D(q_1 \parallel q)$
$D(q_2 \parallel \hat{q}_{12})$
w_2, q_2
$D(\hat{q}_{12} \parallel q)$
$D(q_2 \parallel q)$
q

Figure 3.3: Pictorial representation of the property (3.91) of the D_{FKL}-barycenter of two weighted distributions.

The divergences appearing in (3.91) are pictorially represented in Fig. 3.3. Alternative forms of the identity (3.91) are the following

$$\sum_{i=1}^{2} w_i D_{FKL}(q_i \| q) = \sum_{i=1}^{2} w_i \big(D_{FKL}(q_i \| \hat{q}_{1,2}) + D_{FKL}(\hat{q}_{1,2} \| q)\big), \quad \forall q \in \mathcal{Q}, \tag{3.92}$$

and

$$\sum_{i=1}^{2} w_i \Big(D_{FKL}(q_i \| q) - \big(D_{FKL}(q_i \| \hat{q}_{1,2}) + D_{FKL}(\hat{q}_{1,2} \| q)\big)\Big) = 0, \quad \forall q \in \mathcal{Q}, \tag{3.93}$$

Consider now a weighted triple $\{w_i, q_i\}_{i=1}^3$ and the weighted pair $\{\{(w_1 + w_2), \hat{q}_{1,2}\}, \{w_3, q_3\}\}$, and recall that for distributions in the exponential family the associativity of the D_{FKL}-barycenters holds true, so that

$$\hat{q}_{1,2,3} = \bar{\Phi}_{D_{FKL}}(\{w_i q_i\}_{i=1}^3) = \bar{\Phi}_{D_{FKL}}(\{(w_1 + w_2)\hat{q}_{1,2}, w_3 q_3\}). \tag{3.94}$$

Then, the identity (3.91) can be extended to weighted triples $\{w_i q_i\}_{i=1}^3$:

$$\begin{aligned} w_1 D_{FKL}(q_1\|q) + w_2 D_{FKL}(q_2\|q) + w_3 D_{FKL}(q_3\|q) = \\ = w_1 D_{FKL}(q_1\|\hat{q}_{1,2,3}) + w_2 D_{FKL}(q_2\|\hat{q}_{1,2,3}) + w_3 D_{FKL}(q_3\|\hat{q}_{1,2,3}) + \\ + (w_1 + w_2 + w_3) D_{FKL}(\hat{q}_{1,2,3}\|q), \end{aligned} \tag{3.95}$$

which holds $\forall q \in \mathcal{Q}$. More in general, one can state the following:

Proposition 3.3.4. *Consider a family $\mathcal{Q}$ and a D-measure such that the barycenters of weighted sets of distributions are associative. If, for any pair $\{w_i q_i\}_{i=1}^2$ of weighted distributions in $\mathcal{Q}$, the following property is satisfied,*

$$\begin{aligned} & w_1 D(q_1\|q) + w_2 D(q_2\|q) \\ & \quad = w_1 D(q_1\|\hat{q}_{1,2}) + w_2 D(q_2\|\hat{q}_{1,2}) + (w_1 + w_2) D(\hat{q}_{1,2}\|q), \quad \forall q \in \mathcal{Q}, \end{aligned} \tag{3.96}$$

then, for any triple $\{w_i q_i\}_{i=1}^3$ of weighted distributions in $\mathcal{Q}$, the following holds true

$$\begin{aligned} & w_1 D(q_1\|q) + w_2 D(q_2\|q) + w_3 D(q_3\|q) \\ & \quad = w_1 D(q_1\|\hat{q}_{1,2,3}) + w_2 D(q_2\|\hat{q}_{1,2,3}) + \\ & \quad + w_3 D(q_3\|\hat{q}_{1,2,3}) + (w_1 + w_2 + w_3) D(\hat{q}_{1,2,3}\|q). \end{aligned} \tag{3.97}$$

$\forall q \in \mathcal{Q}$.

The corresponding proof is short and it is reported following.

Proof. Apply the identity (3.96) to the weighted pair $\{(w_1 + w_2)\hat{q}_{1,2}, w_3 q_3\}$, to get, $\forall q \in \mathcal{Q}$,

$$\begin{aligned} & (w_1 + w_2) D(\hat{q}_{1,2}\|q) + w_3 D(q_3\|q) \\ & \quad = (w_1 + w_2) D(\hat{q}_{1,2}\|\hat{q}_{1,2,3}) + w_3 D(q_3\|\hat{q}_{1,2,3}) + (w_1 + w_2 + w_3) D(\hat{q}_{1,2,3}\|q). \end{aligned} \tag{3.98}$$

Replacing the identities

$$(w_1 + w_2) D(\hat{q}_{1,2}\|q) = w_1 D(q_1\|\hat{q}_{1,2}) + w_2 D(q_2\|\hat{q}_{1,2}) - \big(w_1 D(q_1\|q) + w_2 D(q_2\|q)\big), \tag{3.99}$$

and

$$\begin{aligned} & (w_1 + w_2) D(\hat{q}_{1,2}\|\hat{q}_{1,2,3}) \\ & \quad = w_1 D(q_1\|\hat{q}_{1,2}) + w_2 D(q_2\|\hat{q}_{1,2}) - \big(w_1 D(q_1\|\hat{q}_{1,2,3}) + w_2 D(q_2\|\hat{q}_{1,2,3})\big), \end{aligned} \tag{3.100}$$

into (3.98), after straightforward simplifications one gets (3.97). □

Under the barycenter associativity property (3.73), the repeated application of (3.96) allows to extend the properties (3.96) and (3.97) to weighted sets of distributions of any size n. In the D_{FKL}-barycenter case, such a property will allow, in Chapter 4.4, to define all the adaptive reduction theory for mixture models.

3.3.3 Reverse Kullback-Leibler Divergence Barycenter

As done for the D_{FKL}, in Fig.3.4 is reported the D_{RKL}-bar surface. In the Gaussian case, the D_{RKL} also has a closed form for the barycenter computation, that is:

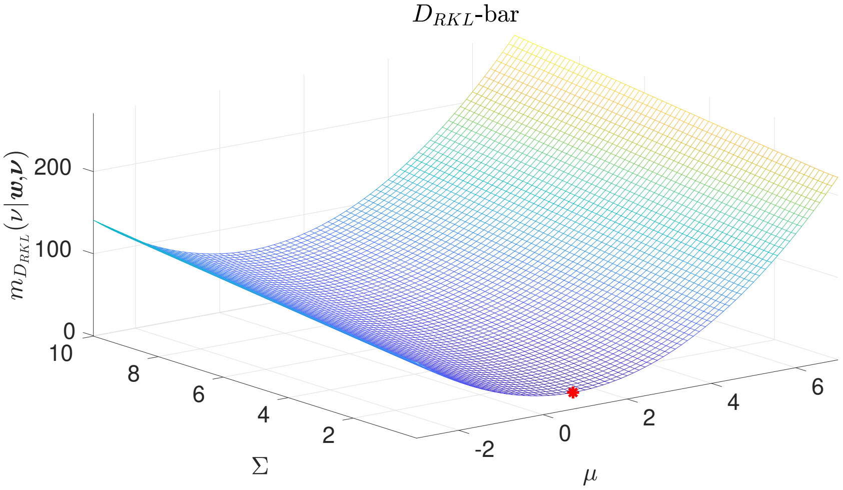

Figure 3.4: D_{RKL}-bar: $\hat{\mu} = 0.7419$, $\widehat{\Sigma} = 0.0968$.

$$\widehat{\Sigma} = \left(\frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \Sigma_i^{-1} \right)^{-1} \triangleq \Psi_{KLA}, \tag{3.101a}$$

$$\hat{\mu} = \widehat{\Sigma} \left(\frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \Sigma_i^{-1} \mu_i \right) \triangleq \mu_{KLA}. \tag{3.101b}$$

which in the literature is known as the Kullback-Leibler Average (KLA) [52] of the Gaussian components. For future purposes, it is also useful to define the following:

$$\tilde{\mu}_{KLA} \triangleq \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \Sigma_i^{-1} \mu_i, \quad \text{so that } \mu_{KLA} = \Psi_{KLA}\, \tilde{\mu}_{KLA}. \tag{3.102}$$

The D_{RKL} is known to be an *exclusive* D-measure, in the sense that it neglects low-density regions in favor of main peaks preservation. If compared to the D_{FKL}, one can notice how narrow the corresponding D_{RKL}-bar covariance is w.r.t. the D_{FKL}-bar one. Moreover, as already mentioned the D_{RKL} exists unique. Regarding the Best Single Gaussian Approximation (BSGA), it does not coincide with the D_{RKL}-barycenter; moreover, the corresponding solution is not unique, and it can incur in numerical issues.

Let us now consider a set of weighted inverse-Wisharts $(\boldsymbol{w}, \boldsymbol{\varphi}) = \{w_i \varphi_i\}_{i=1}^n$; the D_{RKL}-barycenter is then given by:

$$\hat{v} = \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i v_i, \tag{3.103a}$$

$$\widehat{V} = \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i V_i, \tag{3.103b}$$

that is, in the D_{RKL} case, the barycenter of a set of inverse-Wisharts is provided by the weighted arithmetic mean of the parameters.

Given a set of gamma densities $(\boldsymbol{w}, \boldsymbol{\gamma}) = \{w_i \gamma_i\}_{i=1}^n$, the D_{RKL}-barycenter is parameterized as:

$$\hat{\kappa} = \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \kappa_i, \tag{3.104a}$$

$$\hat{\omega} = \left[\frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \frac{1}{\omega_i} \right]^{-1}, \tag{3.104b}$$

that is one obtains the weighted arithmetic mean for the shape parameter $\hat{\kappa}$, whereas the weighted geometric mean for the parameter $\hat{\omega}$.

Until now, for the Gaussian case, it has been possible to write all the barycenters in a closed form; as discussed in [48], the D_{FKL} and the D_{RKL} are the only two cases regarding the reported measures in Sec. 3.1 for which it is possible to obtain analytic solutions. In this regard, all the dissimilarities following in this section do not possess closed forms for the Gaussian case; nonetheless, as mentioned, one can resort to FPI algorithms to obtain an accurate estimate of the corresponding barycenter. As it will be shown, though, some D-measures do not admit a unique minimizer for the barycenter problem. The D_{RKL}-barycenters are also associative, but such a property is not exploited in this dissertation, hence no further analysis will be provided in this regard.

Note: among the listed D-measures, the D_{FKL} and the D_{RKL} are the only two cases of barycenter associativity.

3.3.4 Skew Jeffreys' Divergence Barycenter

Let us consider the skew Jeffreys' divergence (3.15); if the set of weighted Gaussians (3.80) is considered, one obtains the surface for (3.70) reported in Fig.3.5.

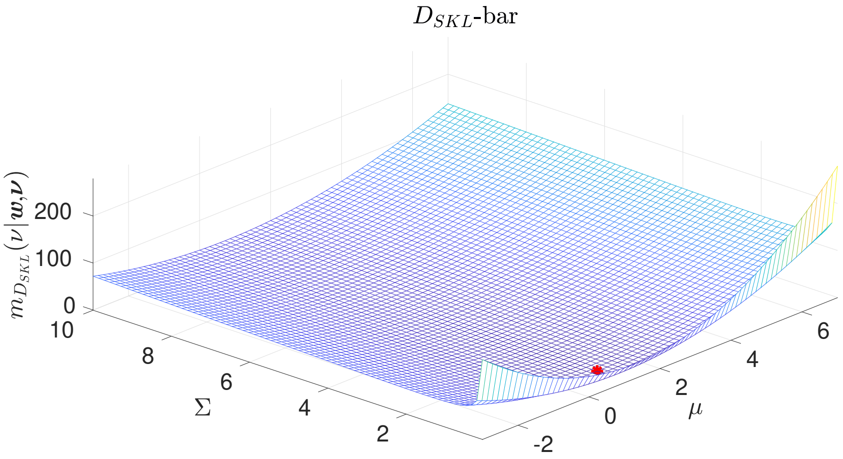

Figure 3.5: D_{SKL}-bar: $\hat{\mu} = 0.8449$, $\widehat{\Sigma} = 0.6624$.

By following the usual approach (3.72), it is not possible to obtain either a closed form solution or a FPI algorithm directly. Nonetheless, by exploiting some algebraic properties, and after some manipulations, it is possible to write the following system of coupled recursive equations:

$$\begin{aligned}\hat{\mu}^{(k+1)} &= (\Gamma^{\alpha,(k)}_{KLA})^{-1}\big[\alpha\tilde{\mu}_{KLA} + (\widehat{\Sigma}^{(}k))^{-1}(1-\alpha)\mu_{MP}\big],\\ \widehat{\Sigma}^{(k+1)} &= (\Gamma^{\alpha,(k)}_{KLA})^{-\frac{1}{2}}\big((\Gamma^{\alpha,(k)}_{KLA})^{\frac{1}{2}}\Gamma^{\alpha,(k)}_{MP}(\Gamma^{\alpha,(k)}_{KLA})^{\frac{1}{2}}\big)^{\frac{1}{2}}(\Gamma^{\alpha,(k)}_{KLA})^{-\frac{1}{2}},\end{aligned} \tag{3.105}$$

where μ_{MP} and $\tilde{\mu}_{KLA}$ are defined in (3.81a) and (3.102), and

$$\begin{aligned} \Gamma_{KLA}^{\alpha,(k)} &= \alpha\Psi_{KLA}^{-1} + (1-\alpha)(\widehat{\Sigma}^{(k)})^{-1}, \\ \Gamma_{MP}^{\alpha,(k)} &= (1-\alpha)\Psi_{MP}(\hat{\mu}^{(k)}) + \alpha\widehat{\Sigma}^{(k)}. \end{aligned} \tag{3.106}$$

with Ψ_{KLA} defined in (3.101a) and $\Psi_{MP}(\hat{\mu})$ defined in (3.81b).

Note that, differently from Ψ_{KLA}, the matrix Ψ_{MP} depends on the current mean $\hat{\mu}^{(k)}$, and therefore must be recomputed at each iteration. If $\alpha = 0.5$, the D_{SKL} case (3.16), the equations (3.105) can be simplified:

$$\begin{aligned} \hat{\mu}^{(k+1)} &= \big[\Psi_{KLA}^{-1} + (\widehat{\Sigma}^{(k)})^{-1}\big]^{-1}\big[\tilde{\mu}_{KLA} + (\widehat{\Sigma}^{(k)})^{-1}\mu_{MP}\big], \\ \widehat{\Sigma}^{(k+1)} &= \Psi_{KLA}^{\frac{1}{2}}\big(\Psi_{KLA}^{-\frac{1}{2}}\Psi_{MP}(\hat{\mu}^{(k)})\Psi_{KLA}^{-\frac{1}{2}}\big)^{\frac{1}{2}}\Psi_{KLA}^{\frac{1}{2}}, \end{aligned} \tag{3.107}$$

which can be efficiently implemented using the Cholesky factorizations of the covariance matrices. Note that the above recursions preserve the symmetry and the positive-definiteness of the covariance matrices. As done in [53], a proof of convergence of the iterations (3.105) is not provided. However, an intensive campaign of numerical tests has shown nice convergence properties of the iterations. Indeed, the D_{SKL}-barycenter of a given set is unique, as proved in [51], and this justifies the nice behavior of the fixed-point recursion provided. A good choice for the initial point $(\hat{\mu}^{(0)}, \widehat{\Sigma}^{(0)})$ is the D_{FKL}-barycenter for $\alpha < 0.5$ and the D_{RKL}-barycenter for $\alpha > 0.5$. Indeed, from the definition (3.15), by varying $\alpha \in [0,1]$, one can *slide* between the D_{FKL} and the D_{RKL} barycenters, passing at the D_{SKL}-barycenter for $\alpha = 0.5$.

With respect to the D_{FKL} and the D_{RKL}, the D_{SKL} seems to fall in between in terms of inclusiveness/exclusiveness.

Regarding the BSGA, it does not coincide with the D_{SKL}-barycenter; moreover, the surface corresponding to the D_{SKL}-BSGA exhibits several local minima, hence the corresponding minimizer is not unique.

3.3.5 Square 2-Wasserstein Distance Barycenter

When the W2 is considered as D-measure to evaluate the barycenter of the set (3.80), one obtains the surface reported in Fig. 3.6.

For this D-measure, the barycenter computation for a set of Gaussian densities is rather particular, since one obtains a closed form for the mean

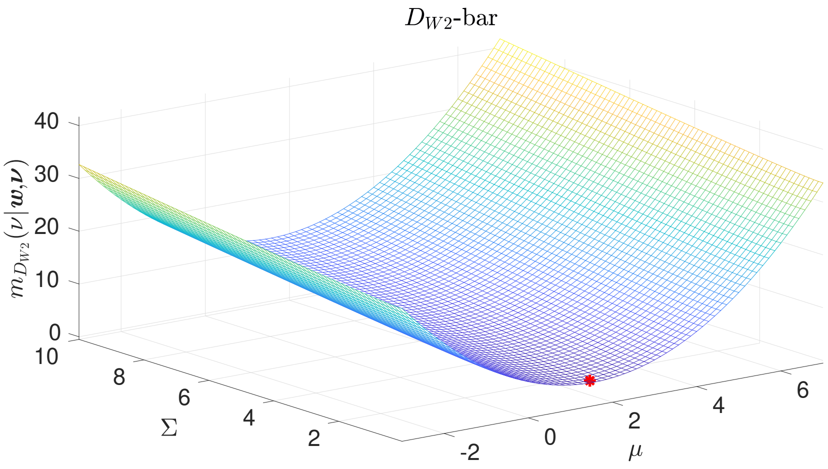

Figure 3.6: D_{W2}-bar: $\hat{\mu} = 1.5500$, $\widehat{\Sigma} = 0.1110$.

and a recursive (FPI) form for the covariance, as:

$$\hat{\mu} = \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \mu_i, \tag{3.108a}$$

$$\widehat{\Sigma}^{(k+1)} = \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \left(\left(\widehat{\Sigma}^{(k)}\right)^{\frac{1}{2}} \Sigma_i \left(\widehat{\Sigma}^{(k)}\right)^{\frac{1}{2}} \right)^{\frac{1}{2}}. \tag{3.108b}$$

Note that the mean (3.108a) of the D_{W2}-barycenter is the μ_{MP} (3.81a), while (3.108b) does not allow a closed form solution for $n > 2$. In [53] the uniqueness of the solution of $\widehat{\Sigma}$ of (3.108b) has been proved. Although no proof of convergence has been provided for the reported form, the authors claim its good convergence properties, which have been verified in an extensive campaign of numerical tests. It is interesting to note that a closed form for the D_{W2}-barycenter covariance exists for $n = 2$ ([54]):

$$\widehat{\Sigma} = \frac{1}{(w_i + w_j)^2} \left(w_i^2 \Sigma_i + w_j^2 \Sigma_j + w_i w_j \left((\Sigma_i \Sigma_j)^{\frac{1}{2}} + (\Sigma_j \Sigma_i)^{\frac{1}{2}} \right) \right). \tag{3.109}$$

Regarding the D_{W2}-BSGA, it is rather difficult to compute (3.69) (even numerically). Nonetheless, an alternative formulation based on the *quantile functions*[5] allows to evaluate numerically the p-th Wasserstein distance; by

[5] Inverse cdfs.

combining such formulation with sampling techniques, it has been found out that the D_{W2}-BSGA coincides with the D_{W2}-barycenter only for the equation relative to the mean, while differing significantly for the covariance solution. The D_{W2}-barycenter appears to be exclusive in terms of covariance, while the D_{W2}-BSGA is particularly inclusive.

All plots reported until now show that, for the D-measures considered so far, suggest the presence of a unique global minimizer (the barycenter); in the works [51,55] is proven the corresponding uniqueness for the general case rather than for $d = 1$ only. It follows that all the proposed FP iterations proposed are quite insensitive to the initialization. As it will be shown, though, for other D-measures this might not be true anymore, and a careful initialization of the FPIs is needed to obtain the convergence to the global minimum of (3.70).

3.3.6 Likeness-based Family Barycenters

As discussed in Sec. 3.1.4, dissimilarity measures such as the *Square L2 norm* (aka *Integral Squared Error (ISE)*) and the *Cauchy-Schwarz divergence (CSD)*, belong to the *Likeness-based family*, which possess closed forms for the dissimilarity between mixtures; jointly to the availability of closed form partial derivatives (3.35), this allows to evaluate both the BSGA and the barycenter quantities. Nonetheless, a main drawback of the LB family is that, even by possessing closed forms for dissimilarities and partial derivatives, the solutions for the problems (3.71) and (3.69) never admit a closed form. For a detailed discussion about gradient-descent computed D_{LB}-BSGAs and D_{LB}-barycenters see [35]. In this work, FPI algorithms ([48]) are instead considered.

Let us define the following quantities:

$$\begin{aligned} \tilde{w}_{i,\nu} &= w_i J^{i,\nu} \ , \quad & \tilde{w}^{c_1}_{i,\nu} &= w_i J^{i,\nu} c_1^{i,\nu}, \quad & \tilde{w}^{c_2}_{i,\nu} &= w_i J^{\nu,\nu} c_2^{i,\nu}, \\ \bar{w}_{\nu} &= \sum_{i=1}^{n} \tilde{w}_{i,\nu} \ , \quad & \bar{w}^{c_1}_{\nu} &= \sum_{i=1}^{n} \tilde{w}^{c_1}_{i,\nu} \ , \quad & \bar{w}^{c_2}_{\nu} &= \sum_{i=1}^{n} \tilde{w}^{c_2}_{i,\nu} \ , \end{aligned} \tag{3.110}$$

where $J^{\cdot,\cdot}$ are the likeness terms as defined in (3.20), and c_i, $i = \{1, 2\}$ are the coefficient proper of each of the LB dissimilarities as discussed in Sec. 3.1.4. It is then possible to derive the following FPI algorithm to compute the Gaussian barycenter $\nu(x|\hat{\mu}, \widehat{\Sigma})$ parameters for a generic D-measure in the

LB family:

$$\hat{\mu}^{(k+1)} = \frac{1}{\bar{w}^{c_1}_{\nu^{(k)}}} \sum_{i=1}^{n} \tilde{w}^{c_1}_{i,\nu^{(k)}} \bar{\mu}_{i,\nu^{(k)}}, \tag{3.111}$$

$$\widehat{\Sigma}^{(k+1)} = \frac{1}{\bar{w}^{c_1}_{\nu^{(k)}}} \Big[\sum_{i=1}^{n} \tilde{w}^{c_1}_{i,\nu^{(k)}} \big(\bar{\Sigma}_{i,\nu^{(k)}} + (\bar{\mu}_{i,\nu^{(k)}} - \hat{\mu}^{(k)})(\bar{\mu}_{i,\nu^{(k)}} - \hat{\mu}^{(k)})^T \big) - \bar{w}^{c_2}_{\nu^{(k)}} \widehat{\Sigma}^{(k)} \Big]. \tag{3.112}$$

where $\nu^{(k)}$ denotes the Gaussian density with parameters $\theta^{(k)} = \{\mu^{(k)}, \Sigma^{(k)}\}$, and $\bar{\mu}$, $\bar{\Sigma}$ are the quantities as defined in (3.26).

Let us now consider the D_{L2} measure when solving the barycenter problem for the set (3.80). The corresponding ADF surface is reported in Fig. 3.7. As anticipated, some D-measures do not admit a unique minimizer for the barycenter problem: the D_{L2} is one of those. Moreover, those multiple solutions could be even equivalent in terms of cost. The existence of several minimizers can represent an issue when addressing the barycenter problem, since the initialization can significantly influence the result. For the D_{L2}, the

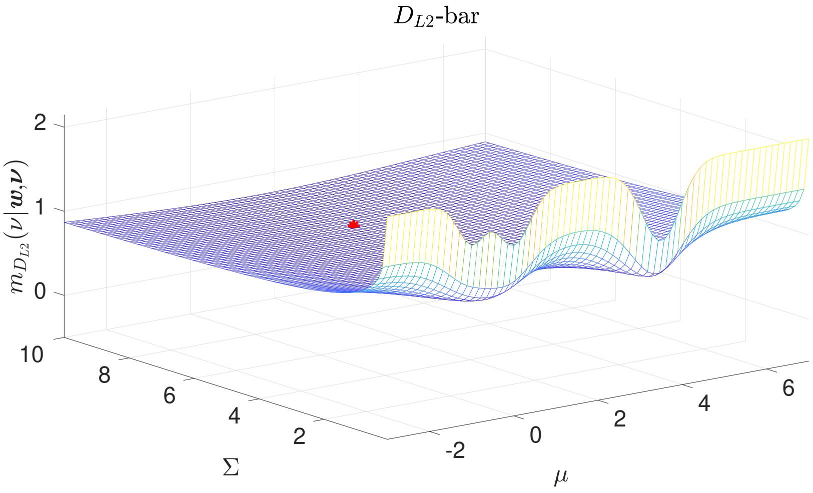

Figure 3.7: D_{L2}-bar: $\hat{\mu} = 1.6105$, $\widehat{\Sigma} = 7.0782$.

general recursion (3.111) becomes:

$$\hat{\mu}^{(k+1)} = \frac{1}{\bar{w}_{\nu^{(k)}}} \sum_{i=1}^{n} \tilde{w}_{i,\nu^{(k)}} \bar{\mu}_{i,\nu^{(k)}}, \tag{3.113}$$

$$\widehat{\Sigma}^{(k+1)} = \frac{1}{\bar{w}_{\nu^{(k)}}} \Big[\sum_{i=1}^{n} \tilde{w}_{i,\nu^{(k)}} \big(\bar{\Sigma}_{i,\nu^{(k)}} + (\bar{\mu}_{i,\nu^{(k)}} - \hat{\mu}^{(k)})(\bar{\mu}_{i,\nu^{(k)}} - \hat{\mu}^{(k)})^T \big) + \tag{3.114}$$

$$+ \frac{1}{2} (\boldsymbol{w}^T \mathbb{1}_n) J_{\nu^{(k)},\nu^{(k)}} \widehat{\Sigma}^{(k)} \Big]. \tag{3.115}$$

When several local minima are present, finding a suitable initialization for such recursion is not an obvious task. The D_{L2} is a rather strange D-measure in terms of features; some observations on its behavior has been discussed even by Runnalls in the work [50]. Sometimes it exhibits particularly exclusive behaviors, which result in pruning-like solutions for the barycenter problem; some others, instead, it becomes a rather inclusive D-measure, as happens in the case reported in Fig.3.7. In general, the lack of solution uniqueness for the barycenter problem is a severe hindrance for a D-measure when applied to the MRP; omitting the details, different mixtures can be equivalent in terms of cost and, moreover, there are no guarantees to reach the global optimum, hence finding often inferior solutions.

Regarding the D_{L2}-BSGA, it coincides with the D_{L2}-barycenter and the corresponding proof is reported in Appendix B.1.2.

Another interesting D-measure in the LB family is the D_{CS} for which one obtains the ADF surface for the set (3.80) as reported in Fig. 3.8. As it will be discussed when considering the Bhattacharyya distance, it is reasonable to assume uniqueness of the solution for the barycenter problem when the Cauchy-Schwarz Divergence is considered. Nonetheless, as for the D_{L2}, the D_{CS} does not admit a closed form for the barycenter problem; nonetheless, one can resort to the following FPI algorithm:

$$\hat{\mu}^{(k+1)} = \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \bar{\mu}_{i,\nu^{(k)}}, \tag{3.116}$$

$$\widehat{\Sigma}^{(k+1)} = \frac{2}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \big(\bar{\Sigma}_{i,\nu^{(k)}} + (\bar{\mu}_{i,\nu^{(k)}} - \hat{\mu}^{(k)})(\bar{\mu}_{i,\nu^{(k)}} - \hat{\mu}^{(k)})^T \big). \tag{3.117}$$

The D_{CS} is a rather strange D-measure itself. If the barycenter is considered, it results to be probably the most inclusive dissimilarity among the ones reported in this work. If the corresponding BSGA, which does not coincide with the barycenter, is instead considered [35], it exhibits a rather counter-intuitive behavior; nonetheless, the corresponding general trend is exclusiveness.

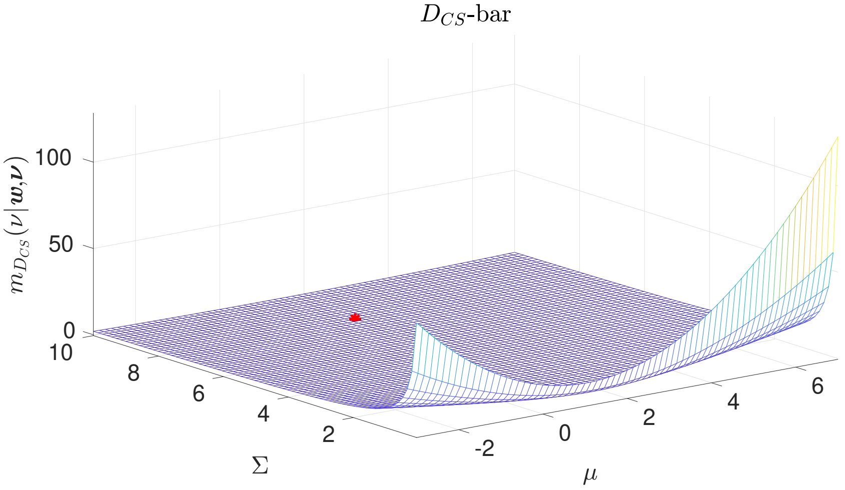

Figure 3.8: D_{CS}-bar: $\hat{\mu} = 1.5407$, $\widehat{\Sigma} = 7.8610$.

3.3.7 I° kind Chernoff α-Divergence barycenter

As done for the LB family, in the following two sections will be discussed the Chernoff α-divergences when applied to the barycenter problem. Let us define the following quantities:

$$w_{i,\nu}^{c_\alpha} = w_i c_\alpha(\nu_i, \nu), \qquad \bar{w}_\nu^{c_\alpha} = \sum_{i=1}^{n} w_{i,\nu}^{c_\alpha}. \tag{3.118}$$

By employing this class of divergences in (3.72), it is possible to obtain the following FPI algorithm:

$$\hat{\mu}^{(k+1)} = \frac{1}{\bar{w}_{\nu^{(k)}}^{c_\alpha}} \sum_{i=1}^{n} w_{i,\nu^{(k)}}^{c_\alpha} \bar{\mu}_{i,\nu^{(k)}}^{\alpha}, \tag{3.119}$$

$$\widehat{\Sigma}^{(k+1)} = \frac{1}{\bar{w}_{\nu^{(k)}}^{c_\alpha}} \sum_{i=1}^{n} w_{i,\nu^{(k)}}^{c_\alpha} \big(\bar{\Sigma}_{i,\nu^{(k)}}^{\alpha} + (\bar{\mu}_{i,\nu^{(k)}}^{\alpha} - \hat{\mu}^{(k)})(\bar{\mu}_{i,\nu^{(k)}}^{\alpha} - \hat{\mu}^{(k)})^T \big), \tag{3.120}$$

where $\bar{\mu}^{\alpha}_{i,\nu}$ and $\bar{\Sigma}^{\alpha}_{i,\nu}$ have been defined in (3.60). As discussed in Sec. 3.1.5, for $\alpha \to 1$ one obtains the D_{FKL}-barycenter, while for $\alpha \to 0$ one should theoretically get the D_{RKL} barycenter in the limit, but the equations as reported here may struggle to converge to such quantity if α is too small. Nonetheless, for $\alpha = 0.5$ one obtains the D_{H2}-barycenter recursive equations. By considering now the D_{H2} in (3.70) for the set (3.80), one obtains the surface reported in Fig.3.9. The D_{H2}, when employed in the barycenter computation, admits several minimizers, and may results to be a particularly exclusive D-measure. If two sharp and well definite peaks are present, the solution corresponding to the barycenter problem is, in general, parameterized as of the two peaks; this is what happens in Fig.3.9. As for the D_{L2} case, the outcome of the barycenter problem is rather sensitive to the initialization. Regarding the D_{H2}-BSGA, it does not coincide with the D_{H2}-barycenter.

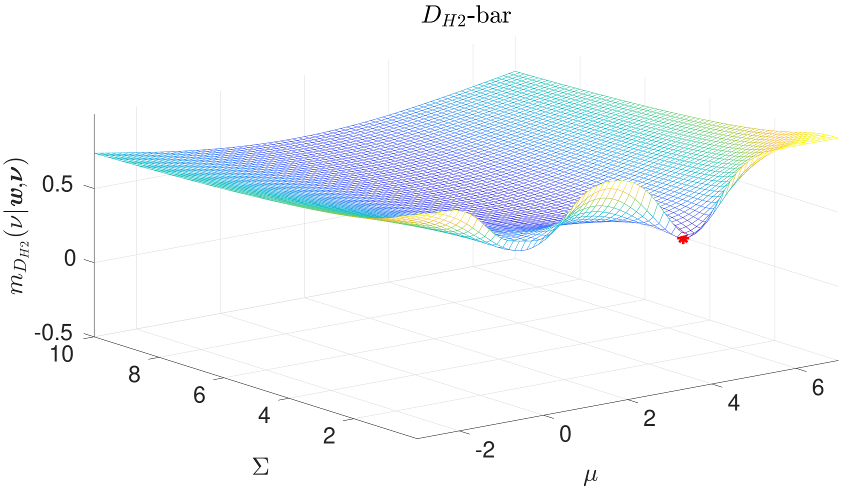

Figure 3.9: D_{H2}-bar: $\hat{\mu} = 3.5000$, $\widehat{\Sigma} = 0.1500$.

3.3.8 II° kind Chernoff α-Divergence barycenter

For completeness, the Chernoff α-divergences of the second kind are analysed as done for all the other D-measures. By employing (A.13) in (3.72), it is possible to obtain the following FPI algorithm:

$$\begin{aligned}\hat{\mu}^{(k+1)} &= \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \bar{\mu}^{\alpha}_{i,\nu^{(k)}}, \\ \widehat{\Sigma}^{(k+1)} &= \frac{1}{\boldsymbol{w}^T \mathbb{1}_n} \sum_{i=1}^{n} w_i \big(\bar{\Sigma}^{\alpha}_{i,\nu^{(k)}} + (\bar{\mu}^{\alpha}_{i,\nu^{(k)}} - \hat{\mu}^{(k)})(\bar{\mu}^{\alpha}_{i,\nu^{(k)}} - \hat{\mu}^{(k)})^T \big).\end{aligned} \tag{3.121}$$

By considering the case $\alpha = 0.5$, one obtains the recursion for the Bhattacharyya distance, which shows again the similarity it bears with the D_{CS} divergence. If the D_B is considered in (3.70) for the set (3.80), one obtains the surface reported in Fig.3.10. The Bhattacharyya distance admits a unique minimizer, hence whatever the initialization, the D_B-barycenter is found by exploiting the reported FPI algorithm. In terms of features, when applied to the barycenter problem, the Bhattacharyya distance is slightly less inclusive than the D_{FKL} measure, favoring peaks more, but still prone to covariance spreading if the components are spread apart in the space. To conclude this section, the D_B-BSGA does not coincide with the D_B-barycenter.

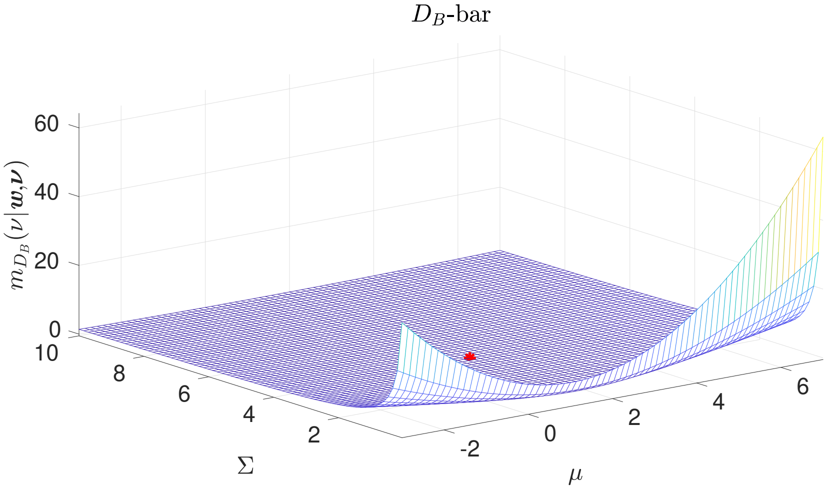

Figure 3.10: D_B-bar: $\hat{\mu} = 1.5316$, $\widehat{\Sigma} = 3.9398$.

3.4 Refinement of reduced order mixture models

Let us assume that a reduced order mixture model is provided, for instance, by means of a greedy reduction algorithm; such a model can serve for a refinement phase where, by exploiting the original mixture information, the corresponding dissimilarity can be further decreased. In the literature have been proposed many algorithms to refine a reduced order model which, in general, can be subdivided in two classes:

- Optimization-based refinement: when the D-measure chosen in (3.68) admits a closed forms between mixtures, one can think to optimize the reduced mixture parameters by means of gradient descent [37]; nonetheless, as discussed, such a property is not common and, among the discussed D-measures in this work, only the LB family is suitable in this regard.

- Clustering-based refinement: as it will be discussed in Chapter 4, when closed forms are not available, it is possible to *induce* composite dissimilarities between mixtures. In general, when this approach is considered, resulting refinement algorithms strongly resemble the K-means and EM algorithms discussed in Chapter 2. Those algorithms will be the core refinement approach in this work, since they can be applied for any D-measure in (3.68).

In both cases, the common goal is to minimize a loss function which can be either the original D-measure or a tractable approximation of it, by exploiting directly the component parameters. Nonetheless, it could often happen that a given tractable approximation might deviate too much from the real D-measure, hence it would be of interest to investigate the features of such alternatives. In this regard, few insights about tractable upper-bounds on the original D-measures will be provided in Chapter 4. For the goals of this work, though, only clustering-based refinement will be reported in detail, since it will serve as last block of a consistent reduction pipeline presented in the next chapter.

3.5 Review of the literature

Until now, many theoretical and fundamental concepts have been discussed without providing any practical mixture reduction algorithm. Nonetheless, all the previously provided blocks are sufficient to describe the majority of the existing approaches; as it will be shown, the general trend when approaching the MRP is to perform a greedy reduction, where a sequence of reduced mixtures decreasing in the size is obtained by either merging or pruning components at each step, and then to pass the corresponding outcome through a refinement phase. Since the most of the new results proposed in this dissertation are concerned with greedy reduction of mixtures, the focus of the literature review will be mainly on such algorithms. Nonetheless, a brief, but rich, bibliography will be reported for the refinement algorithms as well. A greedy reduction algorithm serves in general as a starting point for a refinement phase; in this regard, the author thinks that providing a good initialization is particularly important, both to avoid particularly inferior local minima and to speed up the refinement algorithm convergence.

3.5.1 Greedy reduction algorithms

In order to facilitate the discussion which will follow, it might be convenient to define several terms to further characterize a mixture reduction algorithm. A greedy reduction algorithm is said to be:

- *Global*, if the reduction steps to be performed are aimed at minimizing the dissimilarity between the original mixture and the current reduced order model by taking into account all the mixture parameters.

- *Incrementally global*, if the reduction steps to be performed are aimed at minimizing the dissimilarity between the previous reduced model order and the current one by taking into account all the mixture parameters.

- *Hybrid*, if the reduction performed at each step is aimed at minimizing the dissimilarity between the current approximating mixture and the one reduced by one component by employing only *local* information: the reduction cost is evaluated only as a function of the parameters of pairs of components, and not by considering all the mixture parameters, as it happens for the previous two cases.

- *Local*, if the reduction steps to be performed depend only on the smallest cost of either merging two components, computed simply as the dissimilarity between two components, or pruning one. Such algorithms neglect the mixtures as a whole and focus on the single components.

From a computational perspective, the global algorithms are the most burdensome, while the local approaches are the lightest; nonetheless, the additional cost provides in general a higher approximation accuracy. A fair compromise between accuracy and efficiency is given by the hybrid algorithms.

At the end of each greedy reduction algorithm motivations of consistency, or inconsistency, will be listed.

1989: Kitagawa greedy reduction algorithm (local)

One of the first reduction algorithms for reducing mixture of components has been proposed by Genshiro Kitagawa in the work *Non-Gaussian seasonal adjustment*, where a non-Gaussian state space modeling of time series with trend and seasonality is discussed. In such work, a Gaussian mixture approximation is used to approximate several densities of interest in a Bayesian recursion; nonetheless, such approximations are subject to a rapidly growing number of components, hence making the representations intractable quickly. In this regard, the author proposed a *local* mixture reduction algorithm which, given a mixture of Gaussians $p = \sum_{i=1}^{n} w_i \nu_i$, works as follows:

1. Evaluate the cost of merging two of the mixture components as:

$$D_K(\nu_i \| \nu_j) = w_i w_j D_{SKL}(\nu_i \| \nu_j), \tag{3.122}$$

   where $D_{SKL}(\cdot\|\cdot)$ has been defined in (3.16).

2. Find the indices (i^*, j^*) associated to the pair (ν_i, ν_j) introducing the least value of (3.122):

$$(i^*, j^*) = \underset{i,j}{\arg\min}\, D_K(\nu_i \| \nu_j). \tag{3.123}$$

3. Compute the D_{FKL}-barycenter of the two components and go back to step 1 until a desired number of reduced mixture components has been reached.

Such algorithm results to be really efficient, since local, in terms of computational burden; moreover, from several numerical tests, even the resulting approximations are pretty good. As it will be further discussed, local algorithms tend to introduce a higher degradation in the approximation when many reduction steps are considered.

Inconsistent: merging costs evaluated by means of a heuristic modification of the D_{SKL}, barycenters computed according to the D_{FKL}.

1990: Salmond Joining and Clustering algorithms (local-hybrid)

In the context of target tracking, the mixture reduction problem has been addressed by Salmond [56] who proposed the so called *Joining* and *Clustering* algorithms. In order to discuss such algorithms, let us recall the fact that the D_{FKL}-barycenter of a mixture of Gaussians (3.81) encodes the mean $\hat{\mu}$ and the covariance $\widehat{\Sigma}$ of such a mixture.
Given a Gaussian mixture $p = \sum_{i=1}^{n} w_i \nu_i(x|\mu_i, \Sigma_i)$, the Joining algorithm produces a sequence of reduced order mixtures obtained by greedily merging at each step a pair of components as follows:

1. Evaluate the costs of merging each possible pair of components according to the following dissimilarity measure:

$$D_S(\nu_i \| \nu_j) = \frac{w_i w_j}{w_i + w_j}(\mu_i - \mu_j)^T \widehat{\Sigma}^{-1}(\mu_i - \mu_j), \quad i, j = 1, ..., n, \tag{3.124}$$

   where D_S is a modified version of the Mahalanobis' distance (2.67) and $\widehat{\Sigma}$ is the overall mixture covariance obtained as the D_{FKL}-barycenter of the components.

2. Find the indices (i^*, j^*) associated to the pair (ν_i, ν_j) introducing the least value of (3.124):

$$(i^*, j^*) = \arg\min_{i,j} D_S(\nu_i \| \nu_j). \tag{3.125}$$

3. Compute the D_{FKL}-barycenter of the two components and go back to step 1.

In order to avoid excessive reductions, Salmond proposed an empirical threshold $T = 0.001d$; if the cost associated to the pair of components to be merged falls below such threshold, which represents the maximum acceptable modification of the mixture, and no maximum number of reduced mixture components has been provided, then the merging takes place. If, instead, a maximum number of reduced mixture components is given, and the threshold is violated before reaching such a number, the reduction continues further. A final note on this algorithm is that $\widehat{\Sigma}$ has to be computed just once, since the merging actions are performed by D_{FKL}-barycenter, hence the overall mixture moments are preserved and do not change in the descent. The Joining algorithm can be considered a hybrid algorithm, since the covariance of the whole mixture is involved in the cost computations.

The Clustering algorithm combine components in groups rather than in pairs, but this provides no control over the final reduced mixture number of components. Another difference with the Joining algorithm is given by the assumption that the component associated with the largest weight carries the most of the information. The Clustering algorithm consists in the following steps:

1. Find the component $\nu_c(x|\mu_c, \Sigma_c)$ associated to the largest weight w_c.

2. Given such component, the following cost is evaluated w.r.t. to all the others:
$$\tilde{D}_S(\nu_c \| \nu_i) = \frac{w_i w_c}{w_i + w_c} (\mu_i - \mu_c)^T \Sigma_c^{-1} (\mu_i - \mu_c), \tag{3.126}$$
where w_c, μ_c and Σ_c are the parameters associated to the components with the largest weigth.

3. Merge together all the components for which the cost (3.126) falls below the statistical threshold $T = 0.05 T_1$, where T_1 defines the hyperellipsoid in the space which contains only 1% of the probability mass of the principal component. Go back to step 1.

As discussed, this algorithm does not provide any control over the resulting number of reduced mixture components, since a clustering step might involve more components than the ones necessary to reach the desired mixture size. Moreover, this can be regarded as a purely local algorithm, since it involves only quantities corresponding to single components.

Inconsistent: both algorithms consider merging costs evaluated by means of a heuristic modification of the Mahalanobis' distance, barycenters computed according to the D_{FKL}.

1993: West greedy reduction algorithm (local)

The West algorithm [57] is a local algorithm which, given a mixture p, works as follows:

1. Find the mixture component $\breve{\nu}$ associated to the smallest weight w^*.

2. Evaluate the D_{I2} (ISE) cost from each of the remaining ones.

3. Find the index of the closest component in terms of D_{I2} such that:

$$i^* = \arg\min_i D_{I2}(\breve{\nu} \| \nu_i) < \gamma_w, \tag{3.127}$$

   where $\gamma_w \in \mathbb{R}_+$ is a threshold to prevent the merging of far components. In general $\gamma_w = \infty$ is chosen, which allows the descent to reach a given desired number of components for the reduced order mixture.

4. Merge the pair $(\breve{\nu}, \nu_i^*)$ by means of D_{FKL}-barycenter.

5. Go back to step one until either the desired number of reduced mixture components has been reached or there are no components with a distance smaller than the threshold γ_w.

Similar to the Salmond's Clustering algorithm, the West algorithm only exploits local information in order to reduce the number of mixture components.
Inconsistent: merging costs evaluated according to the D_{I2}, barycenters computed according to the D_{FKL}.

2003: Williams greedy reduction algorithm (incrementally global)

One of the most influential greedy reduction algorithms, in the context of target tracking, is probably that of Williams [37, 58]; such an algorithm is incrementally global according to the D_{I2} (ISE) measure, that is, given a Gaussian mixture $p = \sum_{i=1}^{n} w_i \nu_i$, it works as follows:

1. Evaluate the costs of merging each possible pair of components by means of D_{FKL}-barycenter according to the D_{I2} measure (3.42) between mixtures; since the chosen D-measure is symmetric, one can evaluate half of the merging costs Evaluate the costs of pruning, with subsequent weight renormalization, components of the mixture according to the D_{I2} measure. Each reduction action (either pruning or merging)

generates a mixture which is one order smaller than the preceding one; the corresponding cost is evaluated by computing the D_{L2} between the previous mixture and the one obtained for a given action.

2. Accept as reduced-by-one mixture the one associated with the least ISE w.r.t. the preceding model.

3. Go back to step 1 unless a desired number of components has been reached.

The Williams greedy reduction algorithm results to be rather computationally burdensome. Nonetheless, the author spent a lot of efforts to provide improvements in the implementation.

Inconsistent: merging and pruning costs evaluated by means of D_{L2}, barycenters computed according to the D_{FKL}.

2005: Petrucci greedy reduction algorithm (incrementally global)

A really similar algorithm to the Williams' one is that of Petrucci, where the so called Correlation Measure (CM) is used to proceed in the same way as the Williams reduction algorithm. As discussed in 3.1.4, the correlation measure is equivalent to the D_{CS} in the optimization problem. In this regard, the Petrucci's algorithm can be seen as the Williams' algorithm where the Cauchy-Schwarz Divergence is used instead of the ISE. Even this algorithm is incrementally global, hence computationally burdensome.

Inconsistent: merging and pruning costs evaluated by means of D_{CS}, barycenters computed according to the D_{FKL}.

2007: Runnalls greedy reduction algorithm (hybrid)

The reduction algorithm which appears to be the most interesting one, due to several properties discussed in Chapter 4 in this work, is the Runnalls algorithm [50]. This is a hybrid reduction algorithm, where an upper bound on the real, yet intractable, D_{FKL} between mixtures is evaluated at each reduction step. Given a mixture of Gaussians $p = \sum_{i=1}^{n} w_i \nu_i$, the algorithm works as follows:

1. Evaluate the cost of merging two of the mixture components as:

$$B_{D_{FKL}}(\nu_i \| \nu_j) = \frac{1}{2}\bigg((w_i + w_j) \log |\widehat{\Sigma}_{i,j}| - w_i \log |\Sigma_i| - w_j \log |\Sigma_j| \bigg), \tag{3.128}$$

   where $\widehat{\Sigma}_{i,j}$ is the covariance proper of the D_{FKL}-barycenter (3.81) of the components ν_i and ν_j.

2. Find the indices (i^*, j^*) associated to the pair (ν_i, ν_j) introducing the least value of (3.122):

$$(i^*, j^*) = \arg\min_{i,j} B_{D_{FKL}}(\nu_i \| \nu_j). \tag{3.129}$$

3. Compute the D_{FKL}-barycenter of the two components and go back to step 1 until a desired number of reduced mixture components has been reached.

Aside the relations with the optimal transport framework which will be discussed in Chapter 4, the Runnalls' algorithm is one of the very few algorithms which is totally consistent with a single D-measure, namely the D_{FKL}. This in general yields very good approximations in terms of the intractable D_{FKL} between mixtures. Moreover, being a hybrid algorithm, makes it really efficient in terms of computational resources, with an algorithmic cost comparable with local algorithms previously reported. This is also due to the fact that D_{FKL} barycenters admit a closed form, becoming hence really appealing for real-time applications.

Consistent: merging costs and barycenters computed according to the D_{FKL}.

2010: Enhanced West reduction algorithm (local)

A slight variation to the West algorithm 3.5.1 has been introduced in the work [59], where the authors, given a Gaussian mixture $p = \sum_{i=1}^{n} w_i \nu_i$, suggest the following transformation for the weights:

$$\tilde{w_i} = \frac{w_i}{|\Sigma_i|}, \quad i = 1, ..., n. \tag{3.130}$$

The remainder of the algorithm is equivalent to the standard West algorithm 3.5.1. Although considering such transformed weights can introduce some improvements, the algorithm remains still local, hence not providing accurate approximations.

Inconsistent: merging costs evaluated by means of D_{I2}, barycenters computed according to the D_{FKL}.

2011: Crouse brute-force reduction algorithm (global)

A short survey about reduction algorithms in the context of target tracking has been provided by Crouse [60], where several algorithms are reviewed and a brute-force algorithm is proposed to find the reduced mixture associated to the smallest ISE value. Nonetheless, being a global algorithm evaluating all the possible reduced order models, makes it unsuitable for real-time applications.

Inconsistent: merging costs evaluated by means of D_{I2}, barycenters computed according to the D_{FKL}.

2015: Ardeshiri greedy reduction algorithm (hybrid)

Until now, all the algorithms reported focused only on Gaussian mixtures; nonetheless, as discussed by Ardeshiri in [18], many of the previously reported greedy reduction algorithms can be extended to the exponential family of distributions. During the same year, the author proposed another greedy reduction algorithm [61] based on the D_{RKL} minimization; the interesting perspective there discussed is about the fact that the D_{FKL} measure is a really inclusive D-measure, hence might lead to approximations which assign significant density to regions which, instead, should not contain any. In this regard, considering pruning, given that the D_{RKL} is a rather exclusive D-measure, could avoid such cases to take place.

Inconsistent: merging and pruning costs evaluated by means of D_{RKL}, barycenters computed according to the D_{FKL}.

2018: Square 2-Wasserstein-based greedy reduction algorithm (local)

In the work [54], the authors consider instead the D_{W2} measure to perform the greedy reduction of a Gaussian mixture. The criterion chosen is local and selects for merging, at each reduction step, the two components which are closer in terms of (3.19). Nonetheless, w.r.t. many of the previous algorithms, the merging is performed according to the D_{W2} as in (3.108), hence this algorithm is totally consistent.

Consistent: merging costs and barycenters computed according to the D_{W2}.

2021-2022: Consistent CTD-based greedy reduction (global-hybrid)

As it will be discussed further in the next chapter, in the work [62] has been proposed a general framework for performing greedy reduction by minimizing an induced D-measure between mixtures. Such framework is totally general and can be applied to all the D-measures reported in this work in order to produce consistent greedy reduction algorithms. Moreover, as it will be shown, in some cases such framework allows even to perform a model selection for the reduced mixture which can be embedded in the greedy descent, hence providing a consistent adaptive greedy reduction algorithm.

Consistent: merging costs and barycenters computed according to the same D-measure.

Additional greedy reduction algorithms

Until now, the reported reduction algorithms were mostly methodological discussions to tackle the mixture reduction problem; nonetheless, in the literature some other approaches can be found which deviate substantially from the line followed in this dissertation. In any case, those algorithms might still be of interest in many practical applications. To name a few, one can find the works of Scott [63,64], or the work of Pishdad [65], where a mixture reduction approach is proposed for the target tracking problem which relies, though, on an accurate representation of the process and measurement noises. In the work of [47] another algorithm is proposed to reduce the complexity of a mixture by considering even pruning in the corresponding pipeline.

3.5.2 Refinement algorithms

As discussed in section 3.4, the refinement algorithms rely in general on two main approaches, that is by means of optimization through gradient descent or by optimization by clustering. Since many of the existing approaches exploit complex or sophisticated concepts which would require rather long explanations to be properly discusses, the author of this work will restrict the review to the reporting of those algorithms by trying to enunciate the main principles and potential links with the topics discussed until now.

In 1998, by considering the concept of *virtual samples*, Lippman and Vasconcelos [66] proposed the so called Hierarchical Expectation Maximization (HEM), which can be seen as a *variational*[6] interpretation of the EM algorithm. In 2003, together with the already mentioned greedy reduction algorithm, Williams [37] proposed even a refinement algorithm which performs a gradient descent optimization of the reduced mixture parameters according to the D_{ISE} measure. In 2004 and 2005 respectively, Goldberger [67] and [68] addressed the refinement problem by considering a generalization to the space of distributions of the K-means algorithm, hence by performing a hard-clustering association between the original and reduced mixture components and then by recomputing the reducing mixture components (representatives) as the barycenter of the assigned original components. As done by Williams, Petrucci [41] proposed even a refinement algorithm based on the gradient descent optimization of the reduced mixture parameters according to the CM, which is recalled to be equivalent to the Cauchy-Schwarz divergence. In 2006, Zhang [69] proposed a clustering algorithm based on the D_{ISE} measure; in 2009, Schieferdecker and Huber [70] proposed, after a Runnalls initialization, to perform at first a hard-clustering based on the D_{FKL}, as done by Goldberger and Banjeree, then to perform another clustering, slightly more complex, based on the D_{NISE} between mixtures, then to perform a gradient descent optimization according to the ISE measure and, finally, to perform a weight optimization according, again, to the D_{ISE}. The latter method has been also discussed in [59], after performing an enhanced West algorithm reduction. Nonetheless, in both cases, the weight optimization was not correctly constrained to provide non-negative results, as also pointed out in [60]. However, by recasting such a problem into a quadratic programming problem, it is possible to correctly find the optimized weights, according to the D_{ISE}, for the reduced mixture model. In 2015 and 2016, two additional variational refinement algorithms have been proposed respectively by [71], where a variational upper bound on the Mutual Information (MI) between mixtures is optimized, and by [72], where a particularly tight upper bound on the D_{FKL} between mixtures is minimized. Regarding the latter, a nice

[6] With the term variational are in general denoted approximations for several intractable forms. A good explanation about variational methods in the inference problem can be found in Chapter 10 of [3].

work about D_{FKL} approximations for the Gaussian mixture case has been proposed by [73]. In 2018, an improved version of the HEM algorithm has been proposed by [74], where a very tight bound on the D_{FKL} is optimized in a soft-clustering manner to improve the reduced mixture accuracy. As it will be discussed in the next Chapter, many of the clustering algorithms here reported can be described by the optimal transport theory in a neat way. Nonetheless, as already mentioned, the main goal of this dissertation is to investigate more in detail the greedy reduction, hence initialization, which precedes such refinement algorithms.

3.5.3 Additional approaches

To conclude this section, it is worth to report two more algorithms which are difficult to associate to the previously reported ones. The first one is the Progressive Gaussian Mixture Reduction (PGMR) algorithm, proposed by [75], where a bottom-up approach is followed instead of the classical top-down one. The peculiarity of such algorithm is given by the fact that, by starting from a single component, one performs a split of such density which is optimal, in terms of D_{L2}, w.r.t. the original mixture approximation. Such an algorithm stops when adding more components does not improve anymore significantly the approximation accuracy. The main drawback, though, is given by the fact that in high-dimensional problems it is not obvious nor easy to perform a split of the components. In any case it represents an interesting solution to the MRP. Another algorithm worth to mention is the one proposed by [76], where a variational EM is used to both estimate the reduced mixture parameters, by exploiting directly the original mixture components, and which provides in parallel a suitable number for the reduced mixture components. Providing a suitable number of components is a really uncommon feature for the existing reduction algorithms, yet it represents an impactful choice on the resulting approximation. Of course, if the computational resources force the adoption of a maximum number of components, no matter what features the mixture to be reduced has, one can not do much. Nonetheless, algorithms which provide also a suitable number of components could even warn the user about the suitability of the chosen fixed number of components. As it will be discussed in the remainder of this work, an information about a suitable model order can be very useful in practical problems, even if computational resources are limited.

Chapter 4

Optimal Transport and Consistent Mixture Reduction

4.1 Optimal Transport Theory

In this chapter, the problem of mixture reduction will be seen from another perspective. By exploiting the optimal transport theory, it will be shown how to link hybrid algorithms to a class of induced D-measures in order to perform a consistent mixture reduction. Moreover, the OTT is a general theory which is not restricted to the Gaussian case, hence, all the results here presented can be applied easily to the whole exponential family. In addition, the optimal transport framework works for any problem where the involved mixtures have the same amount of *mass*, which is the case of the problem addressed in this dissertation: this allows to apply the proposed algorithms even to intensities, which represent the uncertainty description in many target-tracking in clutter problems based on the random finite sets (e.g., [24–27] and many more).

The Optimal Transport Problem (OTP) dates back to the year 1781, when Gaspard Monge discussed in the work *Mémoire sur la Théorie des Déblais et des Remblais* the problem of moving optimally a pile of ground from a place to another. Such a problem was concerned with finding a *push-forward* map able to associate uniquely, and with the minimum cost of transportation, infinitesimal amounts of ground between the starting and ending points. The total cost of transporting the whole pile of ground from a place to another is also known in the literature as the Earth Mover's Distance (EMD).

The Monge's formulation can be seen as a geometric perspective of the OTP. Although such a formulation is elegant, it presents several limitations (the problem could be ill-posed in some cases) and often results to be hard to apply in practical problems. A century and a half later (in 1942), Leonid Kantorovich proposed a relaxation of the OTP by working on the problem of moving soldiers between different cities optimally. In this formulation, rather than a transport map, a joint distribution referred to as *coupling* is sought. Given two probability measures[1] $p(x)$ and $q(y)$ in $\mathbb{R}^d$, one seeks a joint distribution $\pi(x, y) \in \Pi(p(x), q(y))$ $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, namely the *coupling* of p and q so that the marginal distributions along x and y coincide with $p(x)$ and $q(y)$. In the Kantorovich problem one has to solve:

$$\mathcal{J} = \inf_{\pi \in \Pi(p,q)} \iint_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \pi(\mathrm{d}x, \mathrm{d}y), \tag{4.1}$$

hence, in this formulation, the OTP can be seen as a loss function which has to be minimized; in this work, the Kantorovich formulation will be used to solve the MRP. By recalling the D_{W2} equation (3.18), it is possible to spot strong similarities; in fact the *Wasserstein metric* is strictly linked with the OTP. The theoretical details are omitted since they would require a lot of argumentation to be introduced; nonetheless, one can find good and intuitive explanations in [77], or in many works of Marco Cuturi, which is a reference figure in this field. For the goals of this work, only the required tools, concepts and ideas will be reported.

4.1.1 Optimal Transport Theory for Mixture Models

As discussed, the OTP deals with the problem of transporting masses from an initial point to a terminal one in a *mass preserving manner* (the mass is neither lost nor gained during the transport) by employing the minimum cost or effort. If one considers probability distributions in a probability space, the OTP is concerned with finding the optimal way of ”mutating” the probability density of one distribution into a target one. When considering mixture models, though, it might not be intuitive to approach such a

[1] In this context, the word measure takes its general mathematical meaning of generalization and formalization of geometrical measures (length, area, volume) and other common notions, such as mass and probability of events. These seemingly distinct concepts have many similarities and can often be treated together in a single mathematical context.

problem. Nonetheless, as discussed in [55], by looking at Gaussian mixtures as discrete probability measures in the space of Gaussian distributions $\mathcal{N}^d$, it is possible to solve the corresponding OTP in an elegant way. Such an extension generalizes to mixtures of any family, and forms the basis for the MRP solution proposed in this dissertation.

4.2 Composite Transportation Dissimilarities

As discussed in Sec. 3.1, in the case of mixtures very few D-measures admit a closed form; this might be a severe limiting factor when addressing the MRP. In this section the concept of *induced* D-measures between mixtures will be defined and discussed regarding the corresponding properties.

4.2.1 Inducing Dissimilarities between Mixtures

Any given D-measure between distributions induces a Composite Transportation Dissimilarity (CTD) defined as described below. For any given pair of mixtures $p(\cdot|\Theta^a)$ and $p(\cdot|\Theta^b)$, of sizes n^a and n^b, respectively, let

$$V(W|\Theta^a, \Theta^b) = \sum_{i=1}^{n^a} \sum_{j=1}^{n^b} W_{i,j} D(q_i^a \| q_j^b), \tag{4.2}$$

where W is a $n^a \times n^b$ matrix in the set $\mathcal{T}(\boldsymbol{w}^a, \boldsymbol{w}^b)$ defined as

$$\mathcal{T}(\boldsymbol{w}^a, \boldsymbol{w}^b) = \{W \in \mathbb{R}_+^{n^a \times n^b} \, : \, W \mathbb{1}_{n^b} = \boldsymbol{w}^a, \; W^T \mathbb{1}_{n^a} = \boldsymbol{w}^b\}. \tag{4.3}$$

(the set of transportation plans, often called *transportation polytope*). Obviously, since $\mathbb{1}_{n^a}^T \boldsymbol{w}^a = \mathbb{1}_{n^b}^T \boldsymbol{w}^b = 1$, it follows that $\mathbb{1}_{n^a}^T W \mathbb{1}_{n^b} = 1$. The Composite Transportation Dissimilarity associated to the given D-measure is defined as

$$C_D(p^a \| p^b) = \min_{W \in \mathcal{T}(\boldsymbol{w}^a, \boldsymbol{w}^b)} V(W|\Theta^a, \Theta^b). \tag{4.4}$$

The minimization problem (4.4) is an Optimal Transport Problem (OTP), a classical Linear Programming (LP) problem, and the variable $W \in \mathcal{T}(\boldsymbol{w}^a, \boldsymbol{w}^b)$ is a *transportation plan* (of the probability masses $\boldsymbol{w}^a$ into $\boldsymbol{w}^b$). Thus, the CTD is the solution of an optimal transport problem. The argument $\widehat{W}$ that minimizes the cost $V(W|\Theta^a, \Theta^b)$ is the *optimal transportation plan.*

The function to be minimized (the cost) can be rewritten in many ways, by defining the cost matrix $\boldsymbol{D}^{a,b} \in \mathbb{R}_+^{n^a \times n^b}$, whose (nonnegative) components are as follows

$$\boldsymbol{D}^{a,b}_{i,j} = D(q_i^a \| q_j^b), \quad i = 1, \ldots, n^a, \quad j = 1, \ldots, n^b, \tag{4.5}$$

one has

$$\begin{aligned} V(W|\Theta^a, \Theta^b) &= \langle \boldsymbol{D}^{a,b}, W \rangle_F \\ &= \operatorname{tr}(W^T \boldsymbol{D}^{a,b}) = \operatorname{vec}(\boldsymbol{D}^{a,b})^T \operatorname{vec}(W). \end{aligned} \tag{4.6}$$

In the previous equation, the symbol $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius scalar product between matrices, $\operatorname{tr}(\cdot)$ is the trace of a matrix, while the $\operatorname{vec}(\cdot)$ operator stacks the columns of a matrix into a single column vector. For the transportation problem the strong duality holds true. The equality constrains of the optimal transport problem

$$W \mathbb{1}_{n^b} = \boldsymbol{w}^a, \qquad \mathbb{1}_{n^a}^T W = (\boldsymbol{w}^b)^T, \tag{4.7}$$

together with the cost function $V(W|\Theta^a, \Theta^b)$ can be easily put in the standard form. Rewriting the identities (4.7) as $I_{n^a} W \mathbb{1}_{n^b} = \boldsymbol{w}^a$ and $\mathbb{1}_{n^a}^T W I_{n^b} = (\boldsymbol{w}^b)^T$ and exploiting the identity $\operatorname{vec}(ABC) = (C^T \otimes A)\operatorname{vec}(B)$, and redefining the optimization variable as $x = \operatorname{vec}(W) \in \mathbb{R}^{n^a \cdot n^b}$, one has

$$\begin{aligned} W \mathbb{1}_{n^b} = \boldsymbol{w}^a \quad &\Longrightarrow \quad (\mathbb{1}_{n^b}^T \otimes I_{n^a}) x = \boldsymbol{w}^a, \\ \mathbb{1}_{n^a}^T W = (\boldsymbol{w}^b)^T \quad &\Longrightarrow \quad (I_{n^b} \otimes \mathbb{1}_{n^a}^T) x = \boldsymbol{w}^b. \end{aligned} \tag{4.8}$$

The primal LP problem (4.4) can therefore can be put in the standard form

$$\begin{aligned} \min \operatorname{vec}(\boldsymbol{D}^{a,b})^T x& \\ \begin{bmatrix} \mathbb{1}_{n^b}^T \otimes I_{n^a} \\ I_{n^b} \otimes \mathbb{1}_{n^a}^T \end{bmatrix} x &= \begin{bmatrix} \boldsymbol{w}^a \\ \boldsymbol{w}^b \end{bmatrix} \\ x &\geq 0. \end{aligned} \tag{4.9}$$

Once such a problem is defined, there exist several solvers which provide a really accurate solution efficiently.

The CTD can then be considered as an *induced* D-measure since, given a base dissimilarity D between components, one can define a dissimilarity between mixtures by solving the problem (4.4). The resulting CTD will inherit the properties of the underlying D-measure (e.g., symmetry) and, often, it results to be an upper bound on the potentially intractable original D-measure between mixtures.

4.2.2 A relaxation of the optimal transport problem

Consider a mixture $p^a = (\boldsymbol{w}^a)^T \boldsymbol{q}^a$ and let $p^b = (\boldsymbol{w}^b)^T \boldsymbol{q}^b$ be one reduced order model of p^a of size $n^b \leq n^a$; moreover, let us define the following problem:

$$\breve{W} = \underset{W \in \breve{\mathcal{T}}(\boldsymbol{w}^a)}{\arg\min} \langle W, \boldsymbol{D}^{a,b} \rangle, \tag{4.10}$$

where:

$$\breve{\mathcal{T}}(\boldsymbol{w}^a) = \{W \in \mathbb{R}_+^{n^a \times n^b} \, : \, W \geq 0, \; W \mathbb{1}_{n^b} = \boldsymbol{w}^a\}, \tag{4.11}$$

is the relaxed transportation polytope, for which it holds that $\breve{\mathcal{T}}(\boldsymbol{w}^a) \supset \mathcal{T}(\boldsymbol{w}^a, \boldsymbol{w}^b)$, and where $\boldsymbol{D}^{a,b}$ is the cost matrix obtained by evaluating the pairwise dissimilarities between the original mixture components and the reduced ones; for instance, if the D_{FKL} is chosen, one has $D_{i,j}^{a,b} = D_{FKL}(q_i^a \| q_j^b)$, that is the pairwise dissimilarity between the i-th original component and the j-th component of the reduced mixture.
The problem (4.10) admits the following closed form solution:

$$\breve{W}_{i,j} = \begin{cases} \frac{w_i^a}{|c(i)|} & \text{if } j \in c(i), \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

where $c(i) = \arg\min_j D_{i,j}^{a,b}$, and $|c(i)|$ is the cardinality of $c(i)$. Given the solution $\breve{W}$ to the relaxed OTP, then it is possible to define the following quantity:

$$C_D^R(p^a \| p^b) = V(\breve{W} | \Theta^a, \Theta^b) = \min_{W \in \breve{\mathcal{T}}(\boldsymbol{w}^a)} \langle W, \boldsymbol{D}^{a,b} \rangle. \tag{4.13}$$

Such a quantity can be considered as an index rather than a dissimilarity between mixtures, since $\boldsymbol{w}^b$ is not preserved, but it will serve as loss function to minimize in a refinement phase.

4.2.3 Joint convexity of D-measures and CTD upper bounds

Many of the reported D-measures in section 3.1 are *jointly* convex [78] in the arguments, that is, given four distributions p_1, p_2, q_1, q_2 and a coefficient $\alpha \in [0, 1]$, it holds that:

$$D(\alpha p_1 + (1-\alpha) p_2 \| \alpha q_1 + (1-\alpha) q_2) \leq \alpha D(p_1 \| q_1) + (1-\alpha) D(p_2 \| q_2). \tag{4.14}$$

In [55,78,79] is reported that the D_{KL}, D_{I2}, D_{W2}, and the general Chernoff α-divergence family are all jointly convex in the arguments; this is not true for many of the D_{LB} measures (aside the D_{I2}). Joint convexity of the underlying D-measure is sufficient to obtain that an induced CTD represents an upper bound for the original D-measure between mixtures, that is given a jointly convex D-measure and two mixture p^a, p^b, it holds that:

$$D(p^a \| p^b) \leq C_D(p^a \| p^b), \quad \forall p^a, p^b \in \mathcal{Q}_{\text{mix}}. \tag{4.15}$$

The proof is reported in Appendix B.2.3.

Properties of CTD in mixture reduction

Why is it important for an induced CTD to be an upper bound for the original D-measure? Often a D-measure between mixtures is intractable, hence having a tractable upper bound allows to solve an approximated MRP in order to find consistent solutions w.r.t. the original D-measure. As it will be shown, the CTD between mixtures can be always computed efficiently if the dissimilarity between single components is available in closed form, and it can be used to formulate both greedy reduction and refinement algorithms. Nonetheless, the CTD can be seen as a dissimilarity measure itself, since, even if it represents a close approximation of the underlying D-measure between mixtures, it may exhibit its own additional features; a nice property of such induced measures is that the CTD-BSDA is equivalent to the D-barycenter, and this can be trivially proven by replacing the CTD formulation in (3.69). Hence, in the CTD framework, the optimal merging action is always the barycenter, which can be efficiently computed if the underlying D-measure admits a single minimizer for the ADF function (3.70) (e.g. by means of FPI algorithms or in a closed form); moreover, in the CTD perspective, pruning is never optimal. As discussed in Section 3.2.3, pruning-like behaviors are strictly linked to the features of a D-measure, and as hinted in the discussion about barycenters, in Section 3.3.1, and it should be something which naturally arises from the properties of a given dissimilarity. From an optimal transport point of view, pruning a component and then performing a weight renormalization is equivalent to split the corresponding mass proportionally to the weights of each of the remaining components, even if significantly dissimilar; clearly, this kind of solution, in the general case, would never be the one obtained by performing an optimal assignment as in the OTP. Alternatively, one could move all the mass towards the most similar component, but that would yield, in general, a higher (or equal) cost if compared to merge both the involved components into their barycenters. The only case where the latter could occur is for measures like the D_{H2}, where it may happen that the barycenter coincides exactly with moving all the mass (weight) of a selected component towards the most similar one, hence simulating a pruning action.

In addition, it can be shown that given an original mixture p^a, the optimal reduced order model p^* is obtained by clustering the original components and by substituting each cluster with its D-barycenter; the CTD as defined in (4.4) exhibits a hard-clustering nature, in the sense that, in any mixture reduction algorithm aimed at minimizing the CTD, the optimally reduced model can be obtained only by clustering together the original components.

Additional notes

In the Kantorovich formulation, the Wasserstein metric arises naturally; nonetheless, the OTP in the case of mixture models can be generalized to any D-measure well-defined for a given family of distributions. In this dissertation, the core investigated D-measure is the D_{FKL}, since, as it will be shown, it possesses sound theoretical properties which make it arguably the best candidate for the MRP as addressed in this work.

4.3 Greedy reduction in the optimal transport framework

The common Mixture Reduction (MR) scheme, composed by a preliminary greedy reduction phase followed by a refinement step, can be formulated using the CTD instead of a plain D-measure. To present the greedy reduction scheme and its properties the following notation will be used: for a given mixture $p^a = (\boldsymbol{w}^a)^T \boldsymbol{q}^a$ let $\hat{q}^a_{i,j}$ denote the D-barycenter of the (unnormalized) submixture $\{w^a_i q^a_i, w^a_j q^a_j\}$, and let $\tilde{p}^a_{i,j}$ denote the *reduced-by-one* mixture obtained from p^a after the replacement of the submixture with the component $(w^a_i + w^a_j)\hat{q}^a_{i,j}$. In formulas:

$$\tilde{p}^a_{i,j} = p^a - \left(w^a_i q^a_i + w^a_j q^a_j\right) + (w^a_i + w^a_j)\hat{q}^a_{i,j}. \tag{4.16}$$

Note: For the sake of discussion, the superscript (m) over a generic mixture p will denote that such mixture contains $m \in \mathbb{N}$ components, which, for simplicity, will also denote the order of the model[2]; in the case of components or weights, the superscript (m) will denote that such quantities belong to the said mixture of order m.

Hybrid C_D-based greedy reduction

As discussed in [62], one can evaluate the cost of merging a pair of components according to the following criterion:

$$B_D(w_i q_i, w_j q_j) = w_i D(q_i \| \hat{q}_{i,j}) + w_j D(q_j \| \hat{q}_{i,j}), \tag{4.17}$$

which can be used to design the reduction scheme reported in Algo. 4.

Algorithm 4: Hybrid C_D-based Greedy reduction Algorithm

```
Data: Original mixture $p^a$, of size $n^a$.
Result: Reduced mixture $p^b$ of size $n^b < n^a$.
$m := n^a$, $p^{(m)} := p^a$ ;
while $m > n^b$ do
    find $(i,j) \in [1\colon m]$:
      $B_D(w_i^{(m)} q_i^{(m)}, w_j^{(m)} q_j^{(m)}) \leq B_D(w_r^{(m)} q_r^{(m)}, w_s^{(m)} q_s^{(m)})$,
      $\forall r > s \in [1\colon m]$;
    $p^{(m-1)} := \tilde{p}_{i,j}^{(m)} = p^{(m)} - w_i^{(m)} q_i^{(m)} - w_j^{(m)} q_j^{(m)} + (w_i^{(m)} + w_j^{(m)}) \hat{q}_{i,j}^{(m)}$;
    $m := m - 1$;
end
$p^b := p^{(m)}$;
```

In this regard, the following theorem represents a core contribution of this dissertation:

Theorem 4.3.1 (B_D and C_D relations)**.** *Given a mixture $p^a = (\boldsymbol{w}^a)^T \boldsymbol{q}^a$ and its reduced-by-one model $\tilde{p}_{i,j}^a$* (4.16)*, it holds that:*

$$C_D(p^a \| \tilde{p}_{i,j}^a) \leq B_D(w_i^a q_i^a, w_j^a q_j^a), \quad \forall i,j \in [1 : n^a], i \neq j. \tag{4.18}$$

[2] In sec. 2.7 the model order was computed by counting all the free parameters; in the mixture case, for the sake of discussion, the corresponding number of components will be regarded as the model order.

If the pair (i^*, j^*) associated to the least value of B_D, namely $\breve{B}_D$, is selected for merging, then

$$C_D(p^a \| \tilde{p}^a_{i^*,j^*}) = B_D(w^a_{i^*} q^a_{i^*}, w^a_{j^*} q^a_{j^*}) = \breve{B}^a_D, \tag{4.19}$$

that is the bound (4.17) *coincides with the C_D between the mixture before and the one after merging the pair (i^*, j^*). Moreover, $\breve{B}_D$ will then denote the minimum increment in CTD.*

All the proofs involved in the above theorem are many and rather lengthy; in this regard, in Appendix B.2.4 is reported all the material. Note that the bound (4.17) is always symmetric, even if the given D-measure is not; this allows to evaluate half of the merging costs at each iteration ($m(m-1)/2$ in total). In [50], for the D_{FKL} case of Gaussian mixtures, it has been proven that (4.17) represents an upper bound on the quantity $D_{FKL}(p^a \| \tilde{p}^a_{i,j})$, which is the analytically intractable D_{FKL} between the mixture before and the one after merging the pair of components minimizing (4.17). In the optimal transport framework, such a bound corresponds to the $C_{D_{FKL}}$ which, given the joint convexity of the D_{FKL}, represents an upper bound for the D_{FKL} between mixtures. In other words, for the D_{FKL} case, the C_D-based greedy reduction algorithm coincides with the Runnalls algorithm 3.5.1. For any other base[3] D-measure, it is possible to obtain a new hybrid greedy reduction algorithm by considering the general algorithmic scheme 4; the D_{W2} case has been investigated in [62].

As it can be noticed, the algorithm provided by the CTD framework is *hybrid*, since it involves only local component parameters in order to infer the increase in dissimilarity between the mixture before and the one after the merging. If the considered C_D represents an upper bound for the original D-measure between mixtures, then such a hybrid algorithm is indirectly optimizing the original, potentially intractable, D-measure. There are though two open issues: the first one is given by the bottleneck of the D-barycenter computation, since, as discussed in Sec. 3.3, only for very few cases one has closed forms; moreover, there might even exist multiple global minima (see D_{I2} or D_{H2}) in the corresponding computation. The second issue is linked with the fact that the resulting C_D might represent a rather crude approximation on the original D-measure; nonetheless, as discussed, the CTD can be considered as a D-measure itself and, if the underlying dissimilarity is

[3] the component-pairwise original D-measure used to induce a CTD.

intractable between mixtures, it still represents a way to approach the MRP for a given D-measure, otherwise intractable. In any case, one should investigate the features of a resulting CTD in order to assess its effectiveness in a greedy reduction scheme. A final note on the greedy reduction algorithm 4 is that it is not restricted to the Gaussian case; if one can compute the D-measure between two generic densities, and can evaluate the corresponding D-barycenter, then Algorithm 4 can be applied to any family of distributions. Moreover, it is not limited to mixture densities, since the barycenter can be computed generically over sets of weighted densities, regardless of their weights; this allows to apply the given algorithm even to intensities (unnormalized mixtures) directly, without requiring a preliminary weight renormalization. The only constraint in the formulation here provided is that the involved mixtures, or intensities, have to share the same amount of mass (sum of the weights). Nonetheless, when working with intensities rather than mixtures, the concept of upper-bound on the original D-measure is lost, since a D-measure as here reported is defined only for distributions. In practice, though, one can still evaluate a dissimilarity between intensities (sharing the same mass) by solving the OTP.

Global C_D-based greedy reduction

Hybrid algorithms as the ones presented before can be suitable for real-time applications if one can compute efficiently the barycenter of components. Nonetheless, they rely on the increase in dissimilarity between contiguous model orders (incrementally global), hence losing the scope over the original mixture during the descent. Given that the CTD can be evaluated in a closed form between mixtures, one could think to formulate global reduction algorithms by computing, at each iteration, the increase in dissimilarity one would introduce w.r.t. the original mixture (and not the previous model order) by merging a couple of components. The corresponding algorithm is reported in Algo. 5.

Basically, this algorithm considers the increase in the C_D introduced w.r.t. the original mixture p^a, as the result of a merging action performed on the current reduced order model. This requires to solve the OTP many times for each model order, hence becoming computationally burdensome. It yields superior approximations if compared to the hybrid algorithm 4, but the resulting gain in the accuracy might not justify the significantly higher computational cost. Nonetheless, if one has a large amount of computational

Algorithm 5: Global C_D-based Greedy reduction Algorithm

```
**Data:** Original mixture $p^a$, of size $n^a$.
**Result:** Reduced mixture $p^b$ of size $n^b < n^a$.
$m := n^a$, $p^{(m)} := p^a$ ;
while $m > n^b$ do
    find $(i,j) \in [1\colon m]$: $C_D(p^a \| \tilde{p}^{(m)}_{i,j}) \leq C_D(p^a \| \tilde{p}^{(m)}_{r,s})$, $\forall r > s \in [1\colon m]$ ;
    $p^{(m-1)} := \tilde{p}^{(m)}_{i,j} = p^{(m)} - (w^{(m)}_i q^{(m)}_i + w^{(m)}_j q^{(m)}_j) + (w^{(m)}_i + w^{(m)}_j)\hat{q}^{(m)}_{i,j}$;
    $m := m - 1$;
end
$p^b := p^{(m)}$;
```

resources available, and no time constraints are presents, using algorithm 5 is suggested if the goal is to achieve better approximations.

As discussed, the CTD framework solves the problem of computing the dissimilarity between mixtures, even if the given D-measure is analytically intractable, and jointly offers a greedy reduction algorithm which can be suitable for real-time applications. As it will be reported in Sec. 4.5, it also provides a general refinement algorithm which can be seen as a generalization of the K-means to the space of distributions. Nonetheless, like most of existing algorithms in the literature, it leaves open the problem of finding a suitable model order for the reduced mixture. Interestingly, as reported in [80], in the D_{KL} case it is instead possible to address such a problem as well with an elegant and efficient criterion. In the next section, a second core contribution of this dissertation is reported; by exploiting the optimal costs $\breve{B}_{D_{FKL}}$, it is possible to figure out when the greedy reduction is introducing a significant dissimilarity w.r.t. the original mixture, hence suggesting which mixture size to halt at the greedy descent.

4.4 Embedding adaptive model selection in the greedy reduction

In the literature, very few mixture reduction algorithms consider the task of finding a suitable number of components for the reduced order model (for instance [75, 76]). Moreover, the existing solutions can be quite computationally burdensome and might not be suitable for real-time applications.

Given a mixture with many components, a reduction algorithm should take into account both the corresponding geometry and the features of the chosen D-measure. Particularly exclusive D-measures would allow in general a higher reduction ratio, since neglecting low density regions would not introduce much of a dissimilarity w.r.t. the original model; in contrast, inclusive D-measures would try to preserve the most of the density, often by assigning significant probability to regions which should not contain any. In applications like target tracking in clutter, the D_{FKL} is often used given its good properties and interpretation (links with Shannon entropy/information or with MLE, or closed form/easy computable barycenters); nonetheless, if there is no active control over the complexity of the reduced order model, it might happen that particularly aggressive reductions could lead to the spread of the component covariances when the involved means are spread apart in the space. This would imply that there is a high probability of finding the state of a target in regions which the original model would totally exclude, increasing the occurrence of filter divergence. Another application where the model complexity plays a central role is unsupervised clustering; given a set of observations, choosing a particularly complex model could lead to overfitting the data, hence reducing the overall generalization ability. In those cases, performing a model simplification could both increase generalization capabilities while providing a more computationally lightweight representation. In both the mentioned applications, it would be of great interest to find a criterion for determining when a given model is suitable for a process of interest, and even capable of deciding that a model is more complex than required, and how much of a reduction one could do before modifying significantly the uncertainty description. Often, in target tracking, the model complexity (number of reduced mixture components) is decided a priori due to computational constrains; nonetheless, it would still be of interest to figure out if such a number can accurately describe the uncertainty or if more components are required. It turns out that, by exploiting again the OTT [80], it is possible to design a lightweight criterion based on the D_{FKL}, which can be used both to characterize the suitability of a model and to provide a potential range of reduction for which a small accuracy is lost in favor of lower computational complexity.

Normalizing the CTD

Let us consider a mixture $p^a = (\boldsymbol{w}^a)^T \boldsymbol{q}^a$, and let $\hat{q}^a$ be its D_{FKL}-barycenter. From a MR perspective, the barycenter represents the coarsest approximation one can provide for a given mixture.
The cost associated to the D_{FKL}-barycenter $\hat{q}^a$ of the mixture p^a is defined as:

$$c(\hat{q}^a|p^a) = \frac{1}{(\boldsymbol{w}^a)^T \mathbb{1}_{n^a}} \sum_{i=1}^{n^a} w_i^a D_{FKL}(q_i^a \| \hat{q}^a) = C_{D_{FKL}}(p^a \| \hat{q}^a). \tag{4.20}$$

from which the next contribution follows: for the D_{FKL} case, by merging components corresponding to the minimum bounds (4.17) as in Algorithm 4, it is possible to obtain a monotonically increasing cost (as function of the decreasing order) which is upper bounded by the mixture barycenter cost $c(\hat{q}^a|p^a)$. This happens thanks to the D_{FKL}-barycenter properties (see Sec. 3.3.2).
Consider a reduced order model $p^{(m)}$ of p^a, $m \leq n^a$, obtained by the $C_{D_{FKL}}$-based greedy reduction algorithm 4; then, by reducing p^a down to its barycenter, it holds that $p^{(n^a)} = p^a$ and $p^{(1)} = \hat{q}^a$. Accordingly, let us define the Relative Transportation Loss (RTL), denoted as $\mathcal{L}$, as:

$$\mathcal{L}^{(m)} = \frac{C_{D_{FKL}}(p^a \| p^{(m)})}{c(\hat{q}^a|p^a)}, \tag{4.21}$$

that is the (global) CTD between the original mixture p^a and its reduced model of order m, divided by the cost of its coarsest possible approximation. For $m = n^a$, one obtains $C_{D_{FKL}}(p^a \| p^a) = 0$, hence $\mathcal{L}^{(n^a)} = 0$, while for $m = 1$, one gets $C_{D_{FKL}}(p^a \| \hat{q}^a) = c(\hat{q}^a|p^a)$, thus $\mathcal{L}^{(1)} = 1$. The overall $\mathcal{L}$ curve will start from zero and will monotonically increase up to one, while the model order decreases from n^a to 1.

Nonetheless, computing $C_{D_{FKL}}(p^a \| p^{(m)})$ for each $m = n^a, ..., 1$ might be computationally burdensome, since the OTP has to be solved for each reduced order model. Luckily, the following property is true:

Proposition 4.4.1. *Consider the sequence of reduced mixtures provided by the hybrid $C_{D_{FKL}}$-based greedy reduction algorithm 4, and consider an index $m < n^a$ (at least one reduction step has been performed). One can define the following function:*

$$\tilde{\mathcal{L}}^{(m)} = \frac{\sum_{l=n^a}^{m} C_{D_{FKL}}(p^{(l)} \| \tilde{p}_{i^*,j^*}^{(l)})}{c(\hat{q}^a|p^a)} = \frac{\sum_{l=n^a}^{m} \breve{B}_{D_{FKL}}^{(l)}}{c(\hat{q}^a|p^a)}, \quad 1 < m < n^a, \tag{4.22}$$

and $\tilde{\mathcal{L}}^{(n^a)} = 0$ *and* $\tilde{\mathcal{L}}^{(1)} = 1$. *Then, the following holds true*

$$\mathcal{L}^{(m)} \leq \tilde{\mathcal{L}}^{(m)}, \quad m \in [1\!:\!n^a]. \tag{4.23}$$

All the discussion regarding summability to one of the RTL is reported in Appendix B.2.5. Since the optimal bounds correspond to the incremental C_D between contiguous model orders (see B.2.4), then even the RTL curve will sum up to one. Such an upper bound corresponds to the cumulative sum of all the l optimal bounds $\breve{B}^{(l)}_{D_{FKL}}$, obtained by merging, for each model order in the descent, the pair of components minimizing the value $B_{D_{FKL}}$. The last equality follows directly from Thm. 4.3.1. Moreover, it holds the following relation:

$$\tilde{\mathcal{L}}^{(m)} \leq s_B^{(m)} \triangleq \frac{\sum_{l=n^a}^{m} B^{(l)}_{D_{FKL}}(w^{(l)}_{i^l} q^{(l)}_{i^l}, w^{(l)}_{j^l} q^{(l)}_{j^l})}{c(\hat{q}^a|p^a)}, \ \forall i^l, j^l \in [1:l], i^l \neq j^l \tag{4.24}$$

where equality always holds for the sequence $(i^{(l^*)}, j^{(l^*)})$ involving the pair of components associated to the optimal merging.

Before proceeding further with the discussion about the adaptive reduction, it is necessary to remark some important concepts; as reported in Theorem 4.3.1, the C_D between two contiguous model orders obtained by optimal merging of components according to (4.17), is equal to the optimal bound $\breve{B}_D$, for any given D-measure. If, instead, a random pair of components are selected for merging, then it holds that $C_D(p^{(m)} \| p^{(m-1)}) \leq B_D(w^{(m)}_i q^{(m)}_i, w^{(m)}_j q^{(m)}_j)$, $\forall i, j \in [1:m], i \neq j$ as in (4.24), that is, in general the bound B_D is greater than the CTD value between a mixture and its reduced-by-one model. In the D_{FKL} case, if non-optimal merging actions are considered, the cumulative sum s_B still sums up to one, but the cumulative sum of the incremental $C_{D_{FKL}}$ terms (the cumulative RTL) may not, since the strict inequality between such quantities may hold. In other words, the summation up to one for both the RTL and the cumulative sum of the $C_{D_{FKL}}$ values happens only if the reduced order models are obtained by means of algorithm 4, that is by sequentially merging pairs of components associated to the optimal bound $\breve{B}_{D_{FKL}}$.

Thus said, from a broad campaign of numerical tests, it resulted that equality between $\mathcal{L}$ and $\tilde{\mathcal{L}}$ is usually verified; the rare cases in which it does not hold are encountered for reduced order models that introduce large $\tilde{\mathcal{L}}^{(m)}$ increments, which are unlikely to be explored by the model selection criterion which will be presented further in this work. Identifying equality conditions would be an important result, since it allows to exactly know, at each reduction step, the total dissimilarity introduced in terms of $C_{D_{FKL}}$ w.r.t. the original mixture. Despite that, the cost (4.22) has some nice features; first of all, it can be computed efficiently as byproduct of the merging costs (4.17), which are directly available in a greedy reduction like the one proposed in Algorithm 4 (hence, the cumulative cost $\tilde{\mathcal{L}}$ is available on the fly and does not require to evaluate all the model orders). Moreover, such cost can be seen as a cumulative loss function monotonically increasing as the model order decreases, which can be interpreted as the accuracy percentage, with respect to the original mixture model, one is losing, in terms of $C_{D_{FKL}}$, if a given mixture is approximated with a reduced order model. Given that the $C_{D_{FKL}}$ represents an upper bound on the original D_{FKL} between mixtures, then the cost (4.22) can be seen as a conservative evaluation of the relative accuracy percentage one is losing in terms of the real Kullback-Leibler divergence between mixture models. Note that a necessary condition for the discussed results is that a given D-measure possesses the associativity of barycenters and the ABTI properties. Among the D-measures reported in this dissertation, only the D_{FKL} and the D_{RKL} possess such a feature; if the associativity does not hold, neither the s_B curve adds up to one, hence making the corresponding D-measure unsuitable for the algorithm proposed in the next section.

An adaptive greedy reduction algorithm

There are several ways to exploit the cost (4.22) for selecting the number of reduced mixture components; one of those could be to provide a percentage threshold denoting the maximum allowed accuracy loss; if during the descent $\tilde{\mathcal{L}}^{(m)}$ becomes larger than the desired threshold value, then the greedy reduction can be halted; another criterion could be to consider the point-wise slope of the resulting cumulative RTL curve (which is exactly equal to the value of each cost $\breve{B}_{D_{FKL}}$) and to halt the reduction if a prescribed value is exceeded. In general, in extreme cases where the merging of two components would introduce a significant error in the approximation, this criterion could prevent such a merge to take place: as discussed, the D_{FKL} is known to spread the covariance of the resulting barycenter if the components to be merged are far in the space (inclusiveness of the D_{FKL}). Hence, such adaptive method could represent an improvement for the resulting greedy reduction algorithm. In algorithm 6 is reported the proposed adaptive greedy reduction algorithm where the halting condition is provided as accuracy loss percentage threshold.

```
Algorithm 6: Adaptive $C_{D_{FKL}}$-based Greedy reduction Algorithm
Data: Original mixture $p^a$, of size $n^a$,
accuracy percentage threshold $\lambda_{\tilde{\mathcal{L}}}$.
Result: Reduced mixture $p^b$ of size $n^b \leq n^a$.
$m := n^a$, $p^{(m)} := p^a$, $\tilde{\mathcal{L}}^{(m)} := 0$;
Compute $c(\hat{q}^a|p^a)$;
while $\tilde{\mathcal{L}}^{(m)} \leq \lambda_{\tilde{\mathcal{L}}}$ do
    find $(i, j) \in [1\colon m]$:
      $B_{D_{FKL}}(w_i^{(m)} q_i^{(m)}, w_j^{(m)} q_j^{(m)}) \leq B_{D_{FKL}}(w_r^{(m)} q_r^{(m)}, w_s^{(m)} q_s^{(m)})$,
      $\forall r > s \in [1 : m]$ ;
    $\tilde{\mathcal{L}}^{(m-1)} := \tilde{\mathcal{L}}^{(m)} + \frac{\breve{B}^{(m)}_{D_{FKL}}}{c(\hat{q}^a|p^a)}$;
    if $\tilde{\mathcal{L}}^{(m-1)} \leq \lambda_{\tilde{\mathcal{L}}}$ then
        $p^{(m-1)} := \tilde{p}_{i,j}^{(m)} = p^{(m)} - w_i^{(m)} q_i^{(m)} - w_j^{(m)} q_j^{(m)} + (w_i^{(m)} + w_j^{(m)}) \hat{q}_{i,j}^{(m)}$;
    end
    $m := m - 1$;
end
$p^b := p^{(m)}$;
```

In order to provide more insights on this approach, let us define the *scaled* D_{FKL} between two mixtures p^a and $p^{(m)}$, denoted $\bar{D}_{FKL}(p^a \| p^{(m)})$, as:

$$\bar{D}_{FKL}(p^a \| p^{(m)}) = \frac{D_{FKL}(p^a \| p^{(m)})}{c(\hat{q}^a|p^a)} \tag{4.25}$$

that is the true D_{FKL} between mixtures divided by the cost of the barycenter $\hat{q}^a$ of p^a. This guarantees that $\bar{D}_{FKL} \leq \mathcal{L} \leq \tilde{\mathcal{L}}$, a condition useful to characterize the $C_{D_{FKL}}$ measure when adopted in greedy reduction procedures.

Unconstrained order mixture reduction

Let us assume that in the greedy reduction phase no desired order n^b has been provided, and that one can let the adaptive algorithm halt the reduction when the chosen condition has been verified. In Fig. 4.1 the result of an adaptive reduction by means of Algorithm 6 for the case of a Gaussian mixture is reported, where the threshold value $\lambda_{\tilde{\mathcal{L}}} = 15\%$ has been considered. By looking at Fig.4.1, many useful insights can be observed; first of all, given the original mixture geometry as the one reported, one could effectively see that the mixture modality can be approximated well with fewer components than $n^a = 40$. With a threshold of 15%, the adaptive algorithms halts the reduction for a model order of $n^b = 10$; the resulting approximation is visually quite accurate. A second observation one can do is that, as mentioned, the real RTL curve, and its cumulative approximation, are equal in this example; although the two curves may not coincide, a broad campaign of numerical experiments showed that in almost all tests the two curves coincide. A third and last observation is that the $C_{D_{FKL}}$ detaches significantly from the real D_{FKL} at some point: from this and many other tests, it has been evinced that the $C_{D_{FKL}}$ represents a tight upper bound for the D_{FKL} between mixtures only for small deviations. In this regard, selecting low values for the RTL threshold could avoid to consider $C_{D_{FKL}}$ values which are considerably distant from the

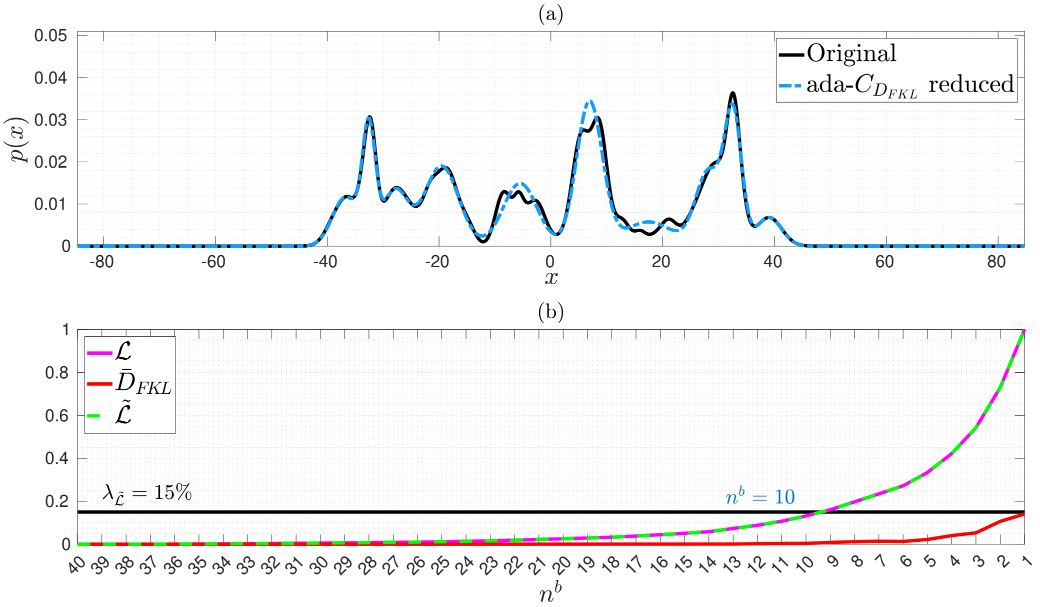

Figure 4.1: (a). Gaussian mixture of size $n^a = 40$ (solid black) is reduced adaptively to $n^b = 10$ components (dashed blue) corresponding to an RTL of 15%. (b). RTL curves (dashed green/magenta), scaled D_{FKL} (solid red).

underlying D_{FKL}; this could serve to obtain reduced order models which are accurate even if the (intractable) D_{FKL} between mixtures is considered.

The experiment has been repeated for the same number of original mixture components, but by spreading the means further in order to obtain more distinct peaks in the model; the resulting reduction is reported in Fig.4.2. As

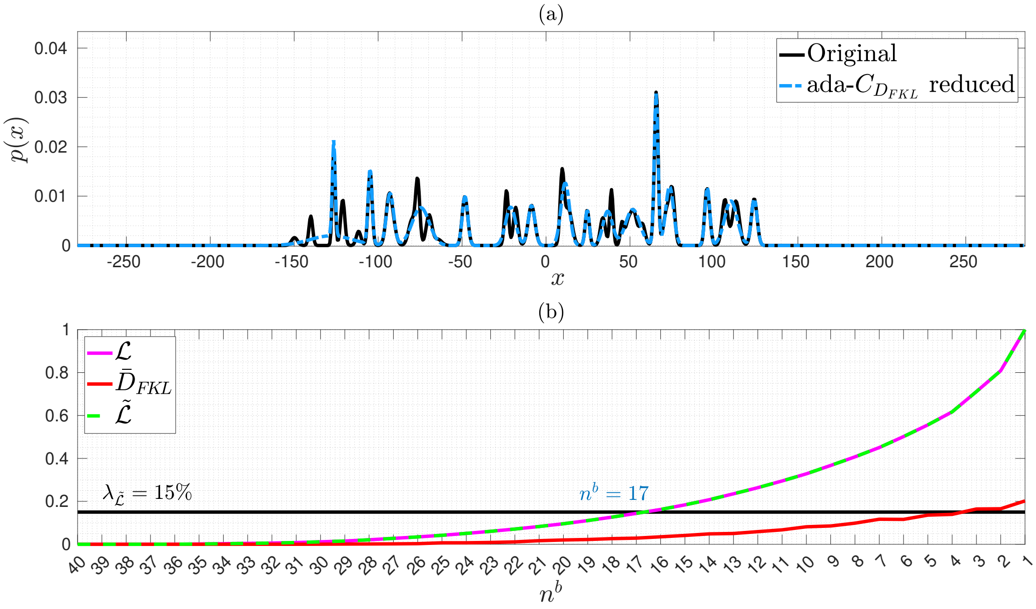

Figure 4.2: (a). Gaussian mixture of size $n^a = 40$ (solid black) is reduced to $n^b = 17$ components (dashed blue) corresponding to an RTL of 15%. (b). RTL curves (dashed green/magenta), scaled D_{FKL} (solid red).

mentioned, the reduced model order should take into account the geometry of the mixture together with the features of the chosen D-measure; in this regard, the proposed method appears to prevent the merging of *far* components, hence halting the reduction earlier if compared to the case reported in Fig. 4.1. Given the inclusiveness of the D_{FKL}, further merging actions would overstretch the covariances of the resulting barycenters, which can be a potentially dangerous effect in contexts like target tracking. W.r.t. Fig. 4.1, in this case are required $n^b = 17$ components rather than $n^b = 10$ in order to loose at most the 15% of the accuracy. Moreover, one can observe that both the RTL curves tend to detach earlier from the horizontal axis if compared to the previous case, hence denoting that the minimum number of reduced mixture components is reached earlier if the same accuracy is wanted to be preserved. Such a feature can be also exploited to figure out if, for a given process, the chosen number of components to represent the uncertainty is adequate; if too small, one can expect the RTL curve to gain a significant slope values already for very low reduction ratios. If too big, one could expect to see the most of the accuracy loss for high reduction ratios. In this regard, let us consider the example reported in Fig. 4.3. This time

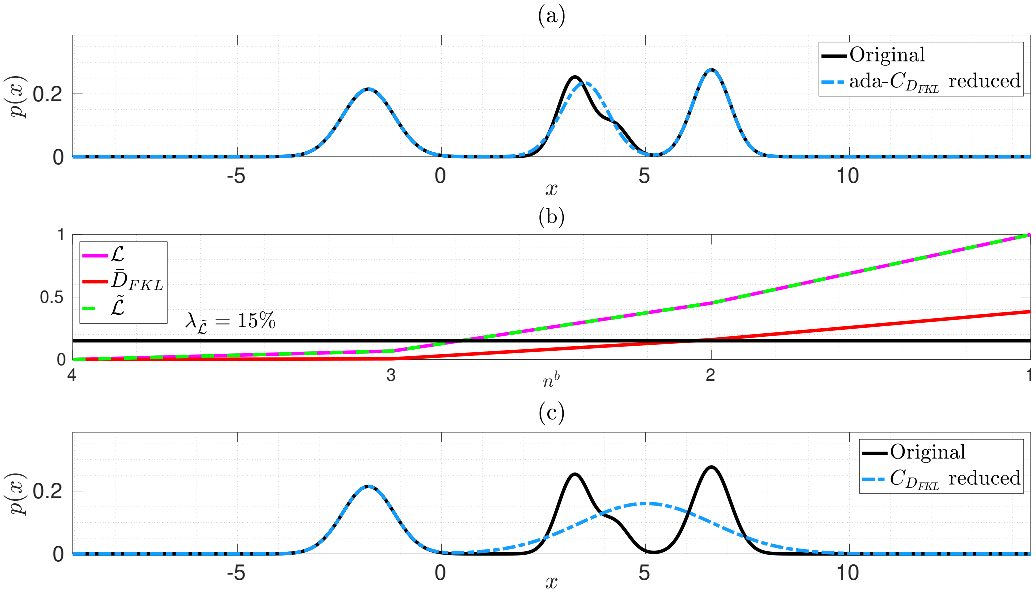

Figure 4.3: (a). Gaussian mixture of size $n^a = 4$ (solid black) is reduced to $n^b = 3$ components (dashed blue) corresponding to an RTL of 15%. (b). RTL curves (dashed green/magenta), scaled D_{FKL} (solid red). (c). Standard greedy reduction down to $n^b = 2$ components (dashed blue).

the original mixture contains $n^a = 4$ components of which two are clustered really close around the value $x = 3.5$. By looking at the corresponding RTL, one can observe that merging the said components introduces a very small dissimilarity, hence it is safe to perform the reduction down to $n^b = 3$ components; nonetheless, if the reduction goes further, a situation like the one reported in Fig. 4.3.(c) arises, that is the subsequent merging tends to assign a significant probability in a region which should not contain any. Moreover, one can see how the RTL curve increases for $n^b = 2$, reaching an RTL value close to 0.5. In contexts like target tracking this may lead to filter divergence.

Relative Transportation Loss halting criteria

Although the proposed criterion seems quite intuitive to apply, there still remains the problem of justifying how to use the RTL curve to halt the reduction. As shown, an approach could be to provide a percentage threshold on the accuracy loss, which, for the sake of discussion, will be denoted *Accuracy Threshold Criterion* (ATC); at this stage, such a choice can rely on several factors, which might include:

- computational capabilities: if the computational resources are limited, then it would be reasonable, in general, to select higher RTL threshold values, allowing a larger accuracy loss, to provide reduced mixtures with a smaller number of components,

- geometry of the mixture: the process of interest generates many components spread apart in the space, hence selecting a too low RTL threshold could quickly push the adaptive number of reduced mixture components towards the maximum allowed amount, if present.

For the conducted experiments, the empirical range of $15\% - 25\%$ represented a good compromise between the accuracy of the resulting approximation and the corresponding number of components. Nonetheless, for the reasons above, the potential user should tune the RTL threshold according to the specific problem of interest.

Another approach could be to consider the slope of the RTL curve, hence to work directly with the normalized costs (4.17) rather than their normalized cumulative sum; for the sake of discussion, such criterion will be denoted *Slope Magnitude Criterion* (SMC). Given the structure of the reduction algorithm 6, if a model can be reduced by several orders, one can expect to find a sequence of costs which have a near-zero value, hence allowing for a modality preserving reduction. In this regard, given a mixture p^a of size n^a, one could consider the criterion of halting the reduction if:

$$\breve{B}^{(m)}_{D_{FKL}} > \alpha \cdot \frac{2}{m} C_{D_{FKL}}(p^{(m)} \| \hat{q}^a), \quad \alpha \in [0, 1]. \tag{4.26}$$

Such an inequality is obtained by inverting B.2.8 in Appendix B.2.5 and by adding a free parameter α; by selecting $\alpha = 1$ one allows for the maximum slope increment, hence the reduction never halts. By instead selecting $\alpha \in (0, 1)$, it is possible to provide an additional halting criterion. Of course, for $\alpha = 0$ no reduction takes place.

Let us define the following quantity:

$$\mathcal{A}^{(m)} = \frac{m}{2} \frac{\breve{B}^{(m)}_{D_{FKL}}}{C_{D_{FKL}}(p^{(m)} \| \hat{q}^a)} \in [0,1] \tag{4.27}$$

obtained by inverting (4.26) and which allows to rewrite such criterion as:

$$\mathcal{A}^{(m)} > \alpha_{\mathcal{A}} \tag{4.28}$$

In section B.2.5 are reported all the proofs regarding such criterion.
The corresponding adaptive algorithm is reported in Algorithm 7.

```
Algorithm 7: Adaptive SMC $C_{D_{FKL}}$-based Greedy reduction Algorithm
Data: Original mixture $p^a$, of size $n^a$,
slope magnitude threshold $\alpha_{\mathcal{A}}$.
Result: Reduced mixture $p^b$ of size $n^b \leq n^a$.
$m := n^a$, $p^{(m)} := p^a$, $\mathcal{A}^{(m)} := 0$;
while $\mathcal{A}^{(m)} \leq \alpha_{\mathcal{A}}$ do
    find $(i,j) \in [1\colon m]$:
      $B_{D_{FKL}}(w_i^{(m)} q_i^{(m)}, w_j^{(m)} q_j^{(m)}) \leq B_{D_{FKL}}(w_r^{(m)} q_r^{(m)}, w_s^{(m)} q_s^{(m)})$,
      $\forall r > s \in [1:m]$;
    $\mathcal{A}^{(m-1)} := \frac{m}{2} \frac{\breve{B}^{(m)}_{D_{FKL}}}{C_{D_{FKL}}(p^{(m)} \| \hat{q}^a)}$;
    if $\mathcal{A}^{(m-1)} \leq \alpha_{\mathcal{A}}$ then
        $p^{(m-1)} := \tilde{p}^{(m)}_{i,j} = p^{(m)} - w_i^{(m)} q_i^{(m)} - w_j^{(m)} q_j^{(m)} + (w_i^{(m)} + w_j^{(m)}) \hat{q}^{(m)}_{i,j}$;
    end
    $m := m - 1$;
end
$p^b := p^{(m)}$;
```

As for the ATC, in order to apply such criterion one has to provide a threshold value $\alpha_{\mathcal{A}} \in [0,1]$. Since the features of $\mathcal{A}$ are pretty different from the ones of $\tilde{\mathcal{L}}$, the choice of such a value is not as intuitive as the case of $\lambda_{\tilde{\mathcal{L}}}$. Nonetheless, from several experiments, the general trend of the SMC is to concentrate the useful reduction range in a very small $\alpha_{\mathcal{A}}$ interval; a value of $\alpha_{\mathcal{A}} = 0.15$[4] seems to provide a good compromise between reduction and accuracy. Values of $\alpha_{\mathcal{A}} = 0.1$ or $\alpha_{\mathcal{A}} = 0.05$ provide really accurate approximations, at the expense of a higher number of components. Of course, for $\alpha_{\mathcal{A}} = 0$ one obtains no reduction, and for $\alpha_{\mathcal{A}} = 1$ the reduction continues down to the barycenter. In this regard, $\alpha_{\mathcal{A}}$ can be used to control the reduction "aggressiveness". In Fig. 4.4 is reported an SMC reduction for values $\alpha_{\mathcal{A}} = \{0.15, 0.1, 0.05\}$ respectively. As it can be observed, by decreasing the slope magnitude threshold one obtains increasingly density-preserving approximations. From now on, in figures, the global RTL curve will be dropped in favor of the new curve $\mathcal{A}$, hence, comparisons between the ATC and the SMC will be provided in parallel to the corresponding curves.

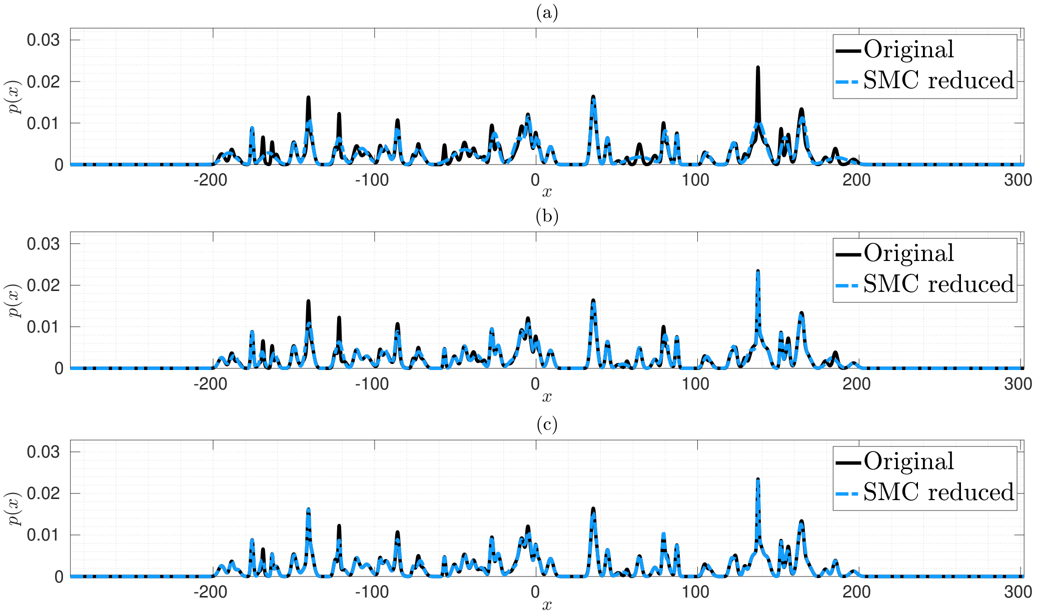

Figure 4.4: (a). Gaussian mixture of size $n^a = 100$ (solid black) is reduced with different slope magnitudes. (a). $\alpha_{\mathcal{A}} = 0.15 \rightarrow n^b = 27$. (b) $\alpha_{\mathcal{A}} = 0.1 \rightarrow n^b = 43$. (c) $\alpha_{\mathcal{A}} = 0.05 \rightarrow n^b = 59$.

To conclude this section, it might be interesting to remark that one could also consider to combine both the discussed criteria together in order to have a guard on the maximum accuracy one is willing to lose. Moreover, if a

[4] From a large campaign of experiments, such a value seems to always provide the best compromise between accuracy and model order.

desired number of reduced mixture components is given (constrained order reduction, discussed in the next section), one should consider to combine all the three halting criteria.

Constrained order mixture reduction

In scenarios where the reduced model order is fixed (say n^b), the previously proposed criteria do not guarantee that the reduced mixture will possess such number of components; nonetheless, to address this constraint, one could consider to let the reduction continue down to n^b, meanwhile computing both the discussed curves. If n^b is reached, and either one of the two thresholds has been exceeded significantly, the user may reconsider, if allowed, what could be a suitable number of components for the process of interest. At the opposite, if n^b is reached, but neither the accuracy threshold or slope magnitude have been violated, one could consider to continue the reduction further, hence saving computational resources, since the desired n^b might be very conservative for the process of interest.

ATC limitations

One limitation of the ATC is represented by the possible reduction interval; for particularly symmetric mixture geometries, or other pathological cases, the RTL curve can tend to a straight line. In such scenarios, the ATC provides deterministically the number of components which will be merged; for instance, given a mixture of size $n^a = 20$, an RTL threshold of 15% would cause the merge of around $\gamma_{\tilde{\mathcal{L}}} \cdot n^a = 3$ components. Nevertheless, such cases are very unlikely to happen in real world problems. Those cases may be associated, in general, to models underfitting the underlying process behavior, where no reduction should actually take place. In addition, given that the RTL curve represents a normalized loss, in the general case, there will always be a short sequence of increasing costs for model orders close to one (to satisfy the summation to one constraint); beside the unlikely case where all the mixture components are equal (singular mixture, for which $c(\hat{q}^a|p^a) = 0$), even for very small deviations the corresponding ATC algorithms would never provide models too low in the order. As it will be proven further in the remainder of this work, such a criterion is in general conservative, often favoring more complex than required models: the SMC criterion can overcome some issues presented by the ATC.

Adaptive reduction halting by accuracy threshold or slope magnitude?

In order to figure out additional features of the two discussed halting criteria, few examples will be reported both to provide further insights on the SMC.

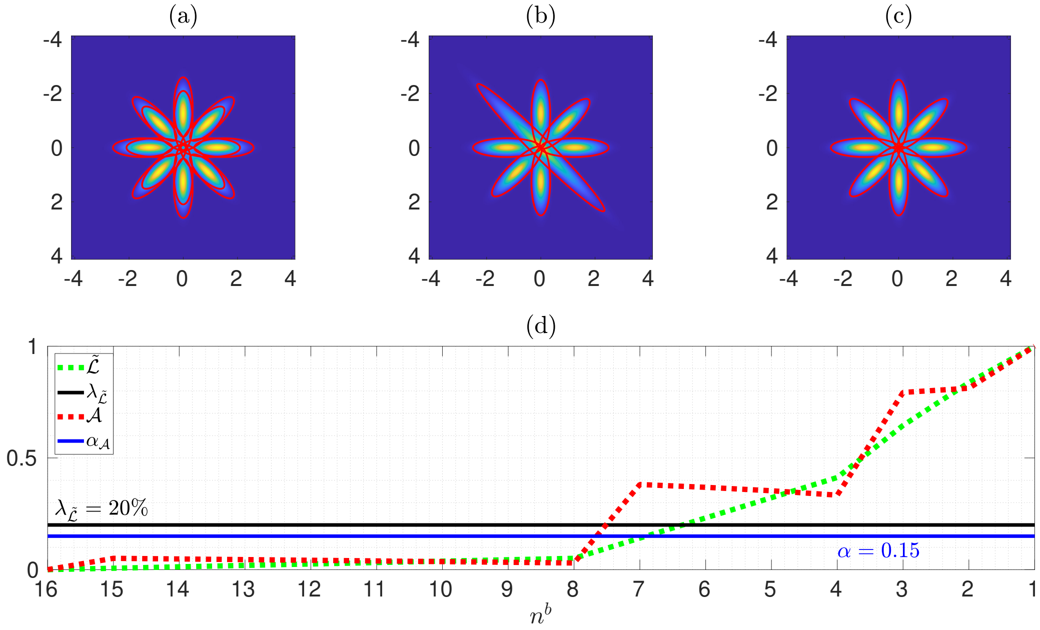

Figure 4.5: (a). Original mixture of size $n^a = 16$. (b). ATC reduced mixture of size $n^b = 7$ for $\lambda_{\tilde{\mathcal{L}}} = 20\%$. (c). SMC reduced mixture of size $n^b = 8$ for $\alpha = 0.15$. (d). Halting criteria.

In Fig. 4.5 is reported a comparison between the two halting criteria for a particularly symmetric mixture of $n^a = 16$ components; the red ellipsoids are the probability curves enclosing 95% of the Gaussian components. As it can be noticed, the original mixture modality could be accurately approximated with only $n^b = 8$ components; in this regard, the SMC, for a value of $\alpha = 0.15$, catches exactly such desired number. Moreover, by looking at both the curves, it is possible to notice a significant slope discontinuity between $n^b = 8$ and $n^b = 7$; this is due to the fact that merging components beyond $n^b = 8$ introduces a high distortion in the mixture modality. If the ATC is considered for $\lambda_{\tilde{\mathcal{L}}} = 20\%$, the reduction halts at $n^b = 7$. Such a behavior often arises when particularly symmetric mixtures are reduced by means of

the accuracy threshold criterion, since it does not provide strict guarantees on the mixture modality preservation. An interesting fact about the $\mathcal{A}$ curve is that, as it can be observed from the figure, if there are several merging actions with the same cost, the corresponding curve tends to decrease over the reduction interval involving those costs. The motivation can be understood by looking at (4.26): the optimal bounds remains constant over several model orders while the cost $C_{D_{FKL}}(p^{(m)}\|\hat{q}^{a})$, and the model order m, slowly decrease; nonetheless, given the optimal choice, the model order decreases faster than the current barycenter cost. From another perspective, once a symmetry/modality is broken, it may have sense to reduce the order further without altering such features significantly. In the above plot, the symmetry can be broken significantly by reducing from $n^b = 8$ to $n^b = 7$ and from $n^b = 4$ to $n^b = 3$.
Another example to be investigated is the benchmark mixture proposed by Crouse [60], which parameters are reported below:

$$\begin{aligned}
\boldsymbol{w} &= \left[0.03, 0.18, 0.12, 0.19, 0.02, 0.16, 0.06, 0.1, 0.08, 0.06\right]^T, \\
\boldsymbol{\mu} &= \left[1.45, 2.20, 0.67, 0.48, 1.49, 0.91, 1.01, 1.42, 2.77, 0.89\right]^T, \\
\boldsymbol{\Sigma} &= \left[0.0487, 0.0305, 0.1171, 0.0174, 0.0295, 0.0102,\right. \\
&\qquad\qquad\qquad \left. 0.0323, 0.0380, 0.0115, 0.0679\right]^T.
\end{aligned} \tag{4.29}$$

In Fig. 4.6 are reported the corresponding reduction outcomes: different behaviors can be observed for the two criteria, and all the previous reasoning can be extended to this case. It is worth to be noted that, by selecting suitably the two thresholds $\lambda_{\tilde{\mathcal{L}}}$ and $\alpha_{\mathcal{A}}$, it may be possible to obtain the same outcome; nonetheless, the two criteria are different in the features and it may be difficult to match the results.

Clustered components case: retrieving the number of clusters

To provide more insights on the effectiveness of the proposed adaptive reduction, let us consider the mixtures generated as following:

1. N_c cluster representatives r_i, $i = 1, ..., N_c$, are generated uniformly in the d-dimensional hypercube of side 2β, $\beta \in [0, \infty)$, that is $r_i \sim \mathcal{U}^d_{[-\beta,\beta]}$; moreover, the corresponding cluster covariances S_i, $i = 1, ..., N_c$, are

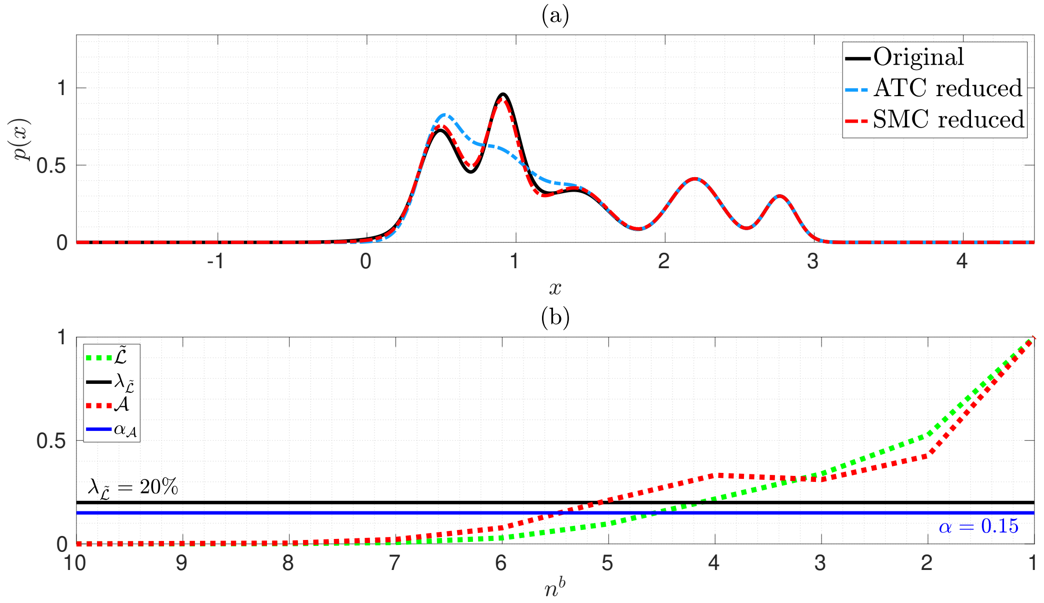

Figure 4.6: (a). Gaussian mixture of size $n^a = 10$ (solid black) is reduced down to $n^b = 5$ components according to the ATC, and to $n^b = 6$ components by means of the SMC. (b) Halting criteria.

generated according to a d-dimensional *Wishart* distribution[5] as $S_i \sim \mathcal{W}(\cdot|\delta_1 I_d, 2d+1)$, $\delta_1 \in [0,\infty)$.

2. Given a number of desired mixture components, say n, a vector of n weights is sampled uniformly in the interval $[0.05, 0.95]$ and then normalized (to obtain $\boldsymbol{w}$); following, a cluster is selected with uniform probability and a mean value is sampled from it as $\mu_i \sim \nu(\cdot|r_i, S_i)$, $i = 1, ..., n$ (to obtain $\boldsymbol{\mu}$). Accordingly, n covariance matrices Σ_i, $i = 1, ..., n$, are sampled as $\Sigma_i \sim \mathcal{W}(\cdot|\delta_2 I_d, 2d+1)$, $\delta_2 \in [0,\infty)$ (to obtain $\boldsymbol{\Sigma}$).

3. A Gaussian mixture $p(x|\Theta) = \boldsymbol{w}^T \boldsymbol{\nu}(x) = \sum_{i=1}^{n} w_i \nu(x|\mu_i, \Sigma_i)$, for which $\Theta = \{\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, is generated.

Such a mixture generation method guarantees, for low values of δ_1, and sufficiently large values of β, to obtain n components grouped in separated clusters. The intent of this experiment is to see how the adaptive algorithms perform in terms of both reducing the model complexity and estimating the number of clusters. In Fig. 4.7 is reported the reduction for an $n^a = 50$ components mixture with both ATC and SMC methods.

In the ATC case, a threshold of 20% amounts to a reduced model of $n^b = 9$ components, hence such criterion misses the underlying number of clusters, but still preserves the modality pretty well. In contrast, the SMC criterion catches exactly the desired number of components ($n^b = 6$) which also correspond to good estimates of the initial clusters. Moreover, in this example something very curious happens; the SMC criterion halts for a value of $n^b = 6$, which amounts for an accuracy loss of around 25%, hence allowing a higher apparent error in favor of lower model complexity. In this regard, if compared to the ATC, the SMC seems to perform better in the problem of finding a good trade-off between model complexity and accuracy. Let us consider now a more complex experiment, reported in Fig. 4.8. The reported experiment does not possess a precisely distinguishable modality: by a visual inspection, the number of clusters could be something between 3 and 4. The ATC, for a value of $\lambda_{\hat{\mathcal{L}}} = 20\%$, yields a reduced mixture of $n^b = 10$ components, which is considerably more than $N_c = 6$. In contrast, the SMC hits again two birds with one bullet, in the sense that not only estimates

[5] The *Wishart* distribution is the distribution defined over the space of Symmetric Positive-Definite (SPD) matrices and can be used to sample covariance matrices for the Gaussian density.

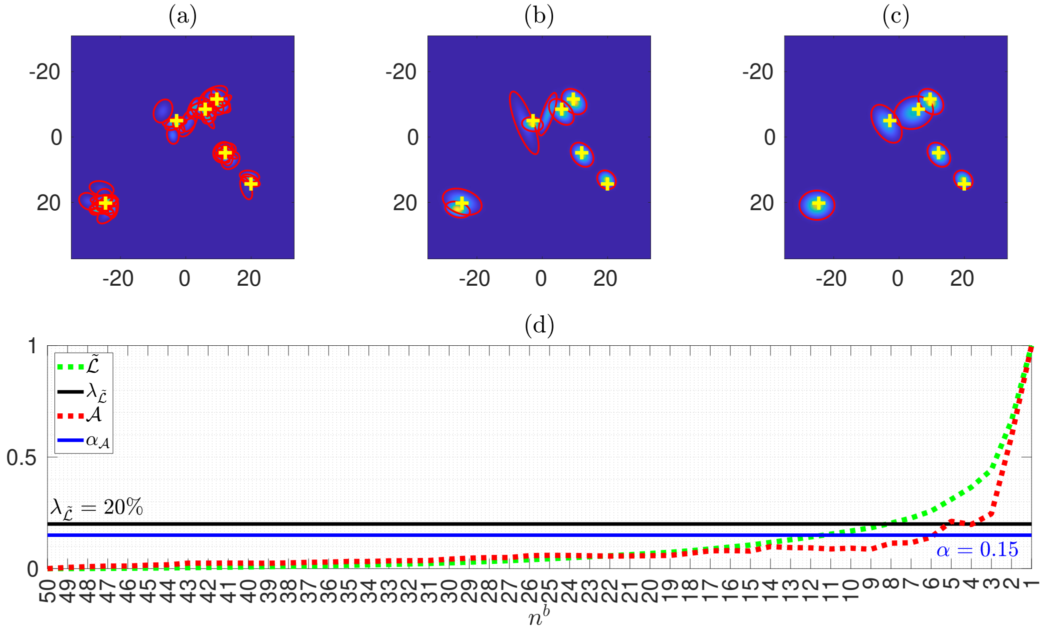

Figure 4.7: (a). Original mixture of size $n^a = 50$ generated out of $N_c = 6$ clusters for parameters $\beta = 25$, $\delta_1 = 0.3$, $\delta_2 = 0.2$; cluster representatives plotted as yellow crosses. (b). ATC reduced mixture of size $n^b = 9$ for $\lambda_{\tilde{\mathcal{L}}} = 20\%$. (c). SMC reduced mixture of size $n^b = 6$ for $\alpha = 0.15$. (d) Halting criteria.

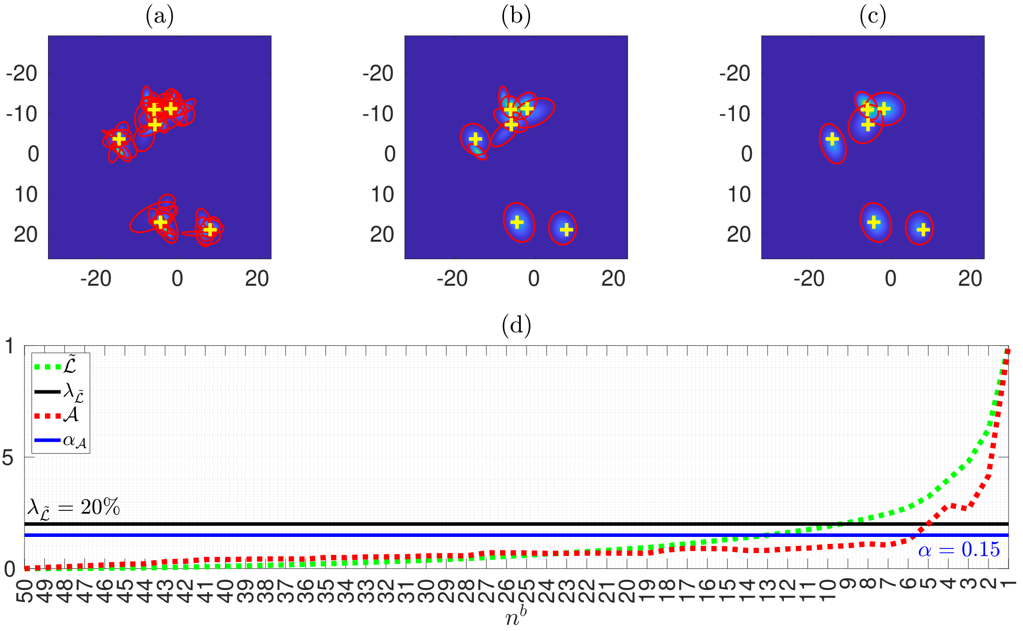

Figure 4.8: (a). Original mixture of size $n^a = 50$ generated out of $N_c = 6$ clusters for parameters $\beta = 25$, $\delta_1 = 0.3$, $\delta_2 = 0.2$; cluster representatives plotted as yellow crosses. (b). ATC reduced mixture of size $n^b = 10$ for $\lambda_{\tilde{\mathcal{L}}} = 20\%$. (c). SMC reduced mixture of size $n^b = 6$ for $\alpha = 0.15$. (d) Halting criteria.

the number of clusters exactly, but it even preserves very well the modality, hence finding a superior trade-off between accuracy and model complexity if compared to the ATC. Even in this case, for a value of $\alpha = 0.15$, the SMC halts the reduction of an accuracy loss higher than the ATC case.

A final remark on the reported experiments is the following: the SMC seems to outperform the ATC in the trade-off between accuracy and model complexity. Nonetheless, it results to be a less intuitive criterion to apply. From several conducted experiments, the best runs regarding SMC performances are achieved for the clustered component cases as the ones reported previously. Such a criterion might be valuable in applications like target tracking in clutter, where object states may be potentially updated with false alarms, hence yielding really similar mixtures as the ones reported in Fig. 4.7 and Fig. 4.8.

For 1-D and 2-D problems it is still possible to visually check the correctness of the model order. But what happens for higher dimensional problems? Let us assume that $n = 60$ components are generated out of $N_c = 10$ clusters, respectively for dimensions $d = 6$ and $d = 12$, and for parameters $\beta = 30$, $\delta_1 = 0.3$ and $\delta_2 = 0.2$; in Fig. 4.9 are reported the corresponding adaptive reductions. Again, the SMC criterion seems to be more versatile than the ATC, even for considerably high dimensional problems.

Additional notes

The adaptive algorithms presented in this section are general and can be applied for any family of mixture distributions or intensities in the exponential family. The curves $\tilde{\mathcal{L}}$ and $\mathcal{A}$, alternatively to undergoing a thresholding in adaptive MR algorithms, can serve as a visual tool for the analysis of suitable model orders, especially in high dimensional problems where such a task becomes non-trivial. As it will be discussed in the next chapter, by coupling the presented adaptive algorithms with clustering solutions like the EM, it is possible to find suitable model orders even for the case of sample data.

4.5 Refinement in the optimal transport framework

To conclude the mixture reduction pipeline, the refinement phase will be discussed. The author wants to remark that the focus of this work has been

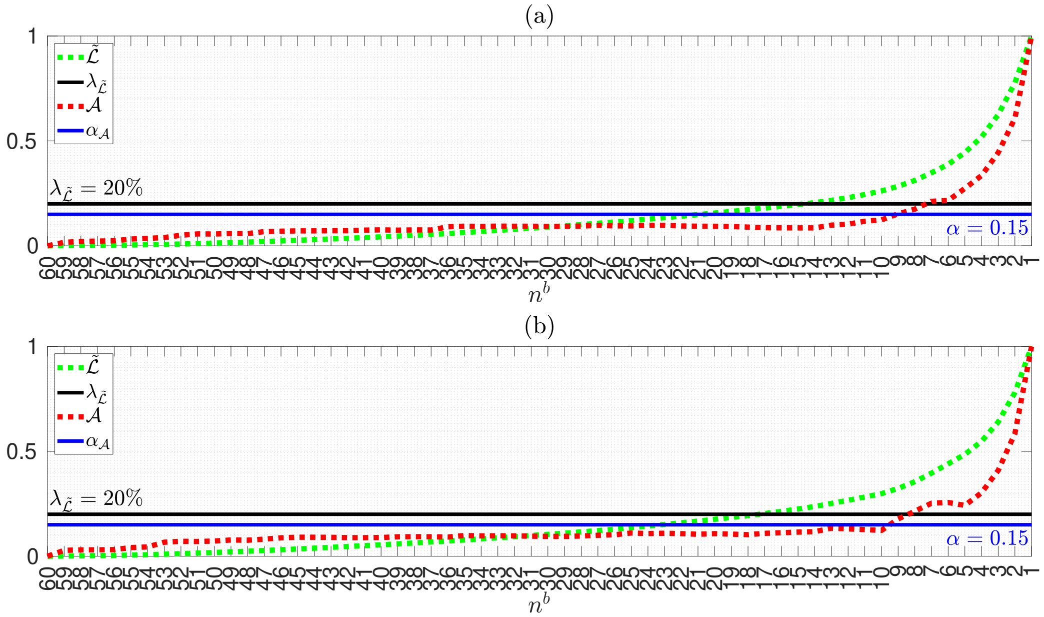

Figure 4.9: (a). $n = 60$ 6-dimensional components are reduced according to the ATC, providing $n^b = 15$ components, and the SMC, yielding $n^b = 10$ components. (b). $n = 60$ 12-dimensional components are reduced according to the ATC, obtaining $n^b = 18$ components, and the SMC, yielding $n^b = 10$ components.

put mostly on the greedy reduction phase, since a superior initialization should provide a faster convergence to better minima in a refinement phase. In Sec. 3.4, a list of existing refinement algorithms has been reported and briefly discussed. In the CTD context, the natural refinement algorithm results to be a generalization of the K-means to the space of distributions; this kind of refinement is something already known in the literature, for instance in the works of [54,67,68,70] the said scheme is presented for several D-measures. Nonetheless, following is reported the corresponding theory visualized in the OTT framework. A discussion regarding the standard K-means has been provided in 2.9, and it should help to understand better the algorithm which will follow.

Assignment by means of relaxed optimal transport

Let us recall the relaxed optimal transport problem (4.13) for which only one constraint is considered. The $\breve{W}$ matrix (4.10) can be seen as a *membership* matrix, which exclusively assigns each of the original components to the reduced mixture representatives.

Update of the reduced mixture components

Recall that $p^a = \sum_{i=1}^{n^a} w_i^a q_i^a$, and let us consider the set $S_{\mathcal{M}} = \{\mathcal{M}_j\}_{j=1}^{n^b}$, where:

$$\mathcal{M}_j = \{\breve{W}_{i,j} q_i^a\}_{i=1}^{n^a}. \tag{4.30}$$

Each $\mathcal{M}_j$ represents a set of weighted densities obtained by (re-)assigning new weights provided by each column of $\breve{W}$, denoted as $\breve{W}_{:,j}$, to the original mixture components. By recalling (3.71), one can update the reduced mixture components by computing the D-barycenter of the assigned original components weighted by each column of the matrix $\breve{W}$ as:

$$\begin{aligned} w_j^b &= \mathbb{1}_{n^a}^T \breve{W}_{:,j}, \\ q_j^b &= \bar{\Phi}_D(\mathcal{M}_j), \end{aligned} \qquad j = 1, ..., n^b, \tag{4.31}$$

that is the j-th reduced mixture component, will be recomputed as the D-barycenter of the set $\mathcal{M}_j$.

As for the K-means algorithm, or more generally according to the Majorization-Minimization (MM) principle, one should iterate between the assignment and update phases in order to minimize the index $\mathcal{J} = \langle \breve{W}, \boldsymbol{D}^{a,b} \rangle$. Let us denote with $p^{(k)}$ the current mixture model at the refinement iteration k, with $\boldsymbol{D}^{a,(k)}$ the cost matrix between p^a and $p^{(k)}$ and with $\breve{W}^{(k)}$ the current relaxed optimal transportation plan computed according to $\boldsymbol{D}^{a,(k)}$; the general C_D-based refinement scheme is reported in algorithm 8. The mixture refinement algorithm here reported results to be particularly efficient for D-measures which possess closed forms both for the pairwise dissimilarity and D-barycenter computation. In the D_{FKL} case, the algorithm 8 coincides with the solutions proposed by [67, 70]. In the D_{W2} case it coincides with the algorithm proposed by [54].

Algorithm 8: C_D-based Refinement Algorithm

```
Data: Original mixture $p^a$ of size $n^a$,
      Reduced mixture $p^b$ of size $n^b$,
      Maximum number of allowed iterations $maxIter$,
      Desired tolerance $tol$.
Result: Refined reduced mixture $p^b$.
$k := 0$, $p^{(k)} := p^b$, $\mathscr{J}^{(k)} := 0$, $\mathscr{J}^{(k-1)} := \infty$;
while $\mathscr{J}^{(k-1)} - \mathscr{J}^{(k)} > tol$ or $k < maxIter$ do
    compute $\boldsymbol{D}^{a,(k)}$;
    compute $\breve{W}^{(k)}$ as in (4.12);
    $\mathscr{J}^{(k-1)} := \mathscr{J}^{(k)}$;
    $k := k + 1$;
    $\mathscr{J}^{(k)} := \langle \breve{W}^{(k)}, \boldsymbol{D}^{a,(k)} \rangle$;
    compute $\boldsymbol{w}^{(k)}$ and $p^{(k)}$ as in (4.31);
end
$p^b := p^{(k)}$;
```

This algorithm completes the consistent mixture reduction scheme obtained by adopting the OTT perspective; from several tests, though, it seems that for the D_{FKL} case, which represents the core D-measure in this work, it rarely introduces a noticeable improvement w.r.t. the greedy reduction result. For any other D-measure, it usually provides significant improvements in terms of CTD.

Hence, which D-measure to use in a mixture reduction?

A final comment on the overall mixture reduction scheme here presented is the following: for each D-measure, different features of the mixture will be preserved in a reduction process. In addition, each D-measure possesses different analytical properties. In this regard, the author wants to recall that the D_{FKL} is the only D-measure for which:

- barycenters are unique and they coincide with the BSDA,
- barycenters possess (semi-)closed forms for the whole exponential family,
- barycenters are associative, hence the MRP solution is affected by less local minima,
- one can perform a theoretically sound adaptive greedy reduction,
- an intuitive interpretation in terms of information can be given / there are direct links with MLE,
- the corresponding CTD represents a good upper-bound on the original, yet intractable, D-measure.

According to the above properties, it is reasonable to consider the D_{FKL} as a very good measure to employ in a mixture reduction problem.

Chapter 5

Numerical Tests and Applications

Most of the material presented in this chapter is rather new and will be soon submitted for publication. All the material following in this chapter represents, for a good part, unpublished contributions.

5.1 Consistency in Mixture Reduction problems

Another topic addressed in this dissertation is that of consistency (alternatively, coherency or congruency) of mixture/intensity reduction pipelines. A preliminary discussion has been proposed in the work [46]; at that time, though, many of the results here presented were not available yet, hence the following discussion is more detailed. In this section, few experiments will be reported to both discuss how consistent reduction pipelines affects the final result in terms of approximation features, and why inconsistent alternatives should be avoided, when possible.

Mixture reduction as optimization problem

In this work, the Mixture Reduction Problem (MRP) (see Chapter 3) has been cast from the beginning as an optimization problem, since this allows to address its solution in a rigorous manner where one aims to minimize a given dissimilarity measure between an original uncertainty representation and a corresponding simplified model. As discussed, though, there are very few cases for which the solution of such a problem can be obtained in a closed form; by exploiting the OTT framework presented in Chapter 4, it is possible to provide many useful tools which allow to address the MRP in an intuitive and efficient manner: if the pairwise dissimilarity D between components can be computed, and the D-barycenter of a set of components can be evaluated either in closed form or by means of FPI algorithms, then it is possible to reduce and refine a mixture coherently with D. Nonetheless, in the OTT framework, one is not optimizing directly the chosen dissimilarity, but a corresponding surrogate[1] function. If such an approximation is sufficiently close to the original D-measure, then reductions done accordingly yield good results even in term of the underlying, yet intractable, measure; for instance, in Sec. 4.4, the $C_{D_{FKL}}$ has been identified as a good approximation of the true D_{FKL} between mixtures. However, few remarks have to be done; first of all, if the induced C_D does not represent an upper bound on the chosen D-measure (or any kind of other approximation), then the corresponding MRP solution may be considerably different; this can happen for measures which are not *jointly convex* in the arguments like, for instance, the Cauchy-Schwarz Divergence (CSD). In [35], a preliminary investigation in this regard has been done, showing that the optimization of the induced $C_{D_{CS}}$ (in that work only the case $n^b = 1$ has been addressed) yields in general very *inclusive* solutions (even more than the D_{FKL}), whereas optimizing the real D_{CS} between mixtures can yield extremely *exclusive* approximations (see the discussion in Section (3.3.1)). In any case, MRPs where the real D-measure is wanted to be optimized result, in general, to be intractable: the OTT offers structured, intuitive and efficient algorithms; on the other hand, the final outcome could be different in terms of preserved features. From now on, the MRP will be addressed in the OTT perspective, where the optimized D-measure is the induced C_D rather than the original, usually intractable, dissimilarity between mixtures.

[1] A surrogate is a function that approximates another function; in general, it is useful because it takes little time to be evaluated. For instance, to search for a point that minimizes a loss function, one could evaluate its surrogate on thousands of points, and take the best value as an approximation to the minimizer of the objective function. In the MR context, the OTT provides surrogate dissimilarity measures between mixtures which can be always evaluated and which, often, represent good approximations on the original dissimilarity.

Why consistent reduction?

As mentioned in Sec. 3.2.1, each D-measure exhibits its own peculiarities; however, there is not much of literature on the corresponding characterization, and it is not clear what dissimilarity measure could be more suitable for a given problem of interest. Performing a mixture reduction, though, can provide some insights in that regard.

In section 3.5 a broad range of the existing algorithms have been reported; a common trend shared by almost all of those is given by the fact that several D-measures are employed in the same reduction pipeline, hence providing incoherent solutions. Why does inconsistency represents a problem in an MRP? As discussed, the problem of reducing the complexity of a mixture can be cast as an optimization problem where a given D-measure is wanted to be minimized; in this regard, considering different dissimilarities in the same pipeline yields, by definition, inferior solutions w.r.t. the problem addressed in a consistent manner. In addition, mixing several D-measures may yield also resulting approximations where heterogeneous features are preserved, even if conflicting each other.

Another common fact in the literature is that the accuracy of an approximation is often evaluated in terms of D_{L2} (D_{ISE}), mostly due to ease of computation. Nonetheless, by considering again the optimization perspective of the MRP, such a fact is equivalent to evaluate how good a solution is in terms of a loss function which may not be involved in the optimization process.

For all the above reasons, the author thinks that preserving consistency in a reduction problem is an important task.

Before proceeding further, let us consider two mixtures $p^a = (\boldsymbol{w}^a)^T \boldsymbol{q}^a$ and $p^b = (\boldsymbol{w}^b)^T \boldsymbol{q}^b$; a list of properties regarding the reported D-measures in Sec. 3.1 follows in Table 5.1.

[2] Closed form only for the case $n = 2$.

***D*-measure properties**				
D-measure	Closed form $D(p^a\|p^b)$	Closed form D-barycenter	$D(p^a\|p^b) \leq C_D(p^a\|p^b)$	D-bar equal to D-BSDA
D_{FKL}	✗	✓	✓	✓
D_{RKL}	✗	✓	✓	✗
D_{SKL}	✗	✗	✓	✗
D_{W2}	✗	✗[2]	✓	✗
D_{L2}	✓	✗	✓	✓
D_{CS}	✓	✗	✗	✗
D_{H2}	✗	✗	✓	✗
D_B	✗	✗	✓	✗

Table 5.1: D-measure properties for each of the reported dissimilarities.

5.1.1 Mixture reduction as a tool of feature analysis for D-measures

For the sake of discussion, the following analysis will be restricted to the Gaussian case, but it remains general.

Let us consider the experiment where an $n^a = 40$ components mixture is reduced according to the algorithm 4, for each D-measure listed in Sec. 3.1; the desired number of components is chosen by at first evaluating an SMC based reduction, which should provide a suitable number of components, and then by diminishing such number by two. The motivation behind this choice relies on the fact that the peculiarities of a D-measure in preserving the mixture features (peaks, support...) become more explicit when a significant information/shape loss is introduced (two less components than those strictly needed). For instance, if the SMC suggests to use $n^b = 6$ components, there will be employed instead $n^b = 4$ Gaussians. Moreover, in order to make the notation lighter, it will be assumed that all the reductions, unless differently stated, will be done according to the Algorithm 4 reported in Sec. 4.3. In this regard, since the reduction algorithm is given, all the resulting plots will be labelled only with the corresponding underlying D-measure. In addition, the original mixture will be shared by each of the reduction algorithms, so even the axis values are removed to save space. In Fig. 5.1 reductions of a $n^a = 40$ GM down to $n^b = 6$ components by employing several D-measures in Algorithm 4 are reported.

The ordering of the plots has been chosen according to the author's ex-

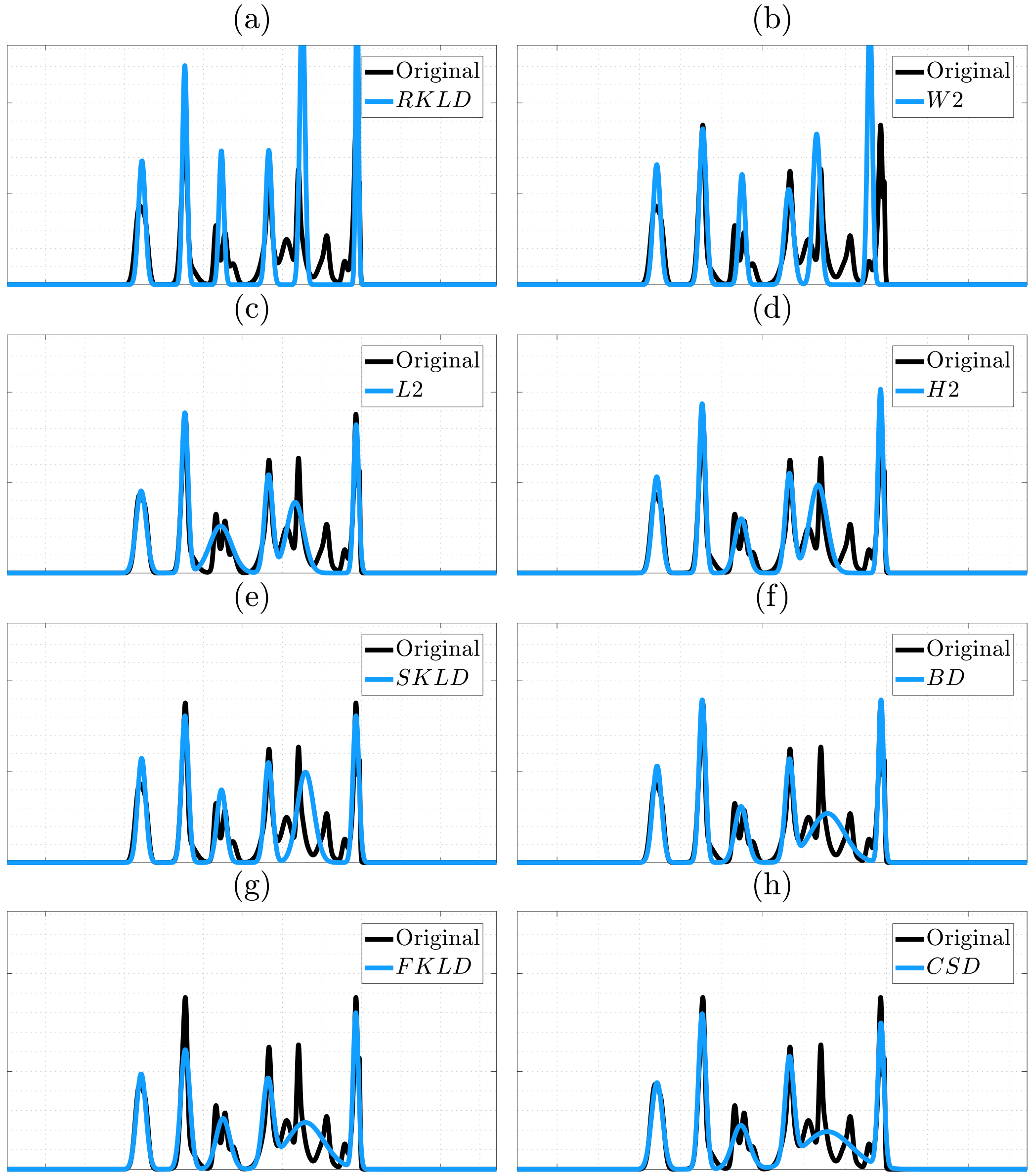

Figure 5.1: GM of size $n^a = 40$ reduced to $n^b = 6$ by consistent C_D-based reduction for several D-measures. (a). $C_{D_{RKL}}$-based reduction. (b). $C_{D_{W2}}$-based reduction. (c). $C_{D_{L2}}$-based reduction. (d). $C_{D_{H2}}$-based reduction. (e). $C_{D_{SKL}}$-based reduction. (f). C_{D_B}-based reduction. (g). $C_{D_{FKL}}$-based reduction. (h). $C_{D_{CS}}$-based reduction.

perience in order to associate similar D-measures in terms of features. By looking at Fig. 5.1, it is possible to provide some preliminary observations. In a C_D-based mixture reduction:

- The D_{RKL} appears to be the most exclusive measure since it preserves only the 6 main peaks, neglecting significant non-zero density regions.

- The D_{W2} results to be comparable with the D_{RKL} in terms of exclusiveness, but this seems to occur mostly for low dimensional problems; from several tests, such a D-measure preserves in general the *geometry* of the mixture (as also discussed in [54]), that is it aims to preserve the modality of the principal modality of the distribution.

- The D_{L2} is, instead, a puzzling D-measure (as also suggested in [50]). From several numerical tests, it can allow very inclusive approximations (by spreading the covariance of the resulting barycenter) over low density regions, but also it can exhibit pruning-like behaviors (for instance, when two "distant" pronounced local minima exist and one is chosen as final approximation), hence by yielding practically a deletion of one component. By recalling Fig. 3.7, the latter argumentation should be more clear. In that figure, the global minima is represented by a very inclusive solution; nonetheless, two local minima, corresponding to pruning one of the two involved components, are possible solutions in some cases. It seems that for sharp peaks, the D_{L2} favors pruning behaviors, while it favors merging with covariance spreading for low importance regions.

- The D_{H2} is an intriguing D-measure to consider in a C_D-based reduction, even if it lacks uniqueness of barycenters. It represents a good trade-off between exclusiveness and inclusiveness, yielding visually interesting approximations, at least in low dimensional problems. As for the D_{L2}, the D_{H2} can exhibit pruning-like behaviors (see Fig. 3.9) if the FPI algorithm is initialized near the main peaks).

- The D_{SKL} falls in between the D_{RKL} and the D_{FKL}; in fact, by looking at Fig. 5.1, it is possible to spot how the main peaks are preserved in a coarser if compared to the D_{RKL}, in favor of assigning non-zero density to less important components as happens for the D_{FKL}.

- The D_B is also a really interesting D-measure; it leads to approximations which are similar to the D_{FKL} ones, but it favors slightly more the main peaks preservation rather than inclusiveness. In general, it yields visually good approximations.

- The D_{FKL} has been investigated extensively in Chapter 4 and in the literature (see [19, 50, 51, 62, 73, 80]), since it represents a golden standard among the statistical divergences; in the context of Fig. 5.1 or, more in general, in the MRP, it is possible to frame it between the D_B and the D_{CS} in terms of inclusiveness: the general trend is a more accentuated spreading of the covariances if compared to the D_B, but not as pronounced as for the D_{CS} for low importance regions.

- The D_{CS} is a rather puzzling D-measure also (as for the whole LB family), since analytically it is very close to the Bhattacharyya distance, hence trying to approximate the peaks, but, at the same time, it favors a spreading of the covariances by around a factor two if compared to the D_{FKL} and the D_B. A clear general trend for this D-measure has not identified yet, similarly to the D_{I2} case, but it has been observed that, in order to preserve the main peaks similarly to the D_B, it favors a very inclusive solution for low density regions.

In Fig. 5.2 is reported another reduction; this time $n^b = 4$. By looking at such figure, it is possible to find back the previously observed features: the D_{RKL} is very exclusive, the D_{W2} follows, the D_{I2} again shows *coexistence of opposite behaviors*, that is it maps tightly the three main peaks, and then spreads the covariance all over the mixture support in order to assign non-zero density to the remaining components. Following, it is possible to observe how the D_{H2} maps the modality of the mixture, but this time less aggressively if compared to the D_{SKL}. D_B, D_{FKL} and D_{CS} again show an increasing factor of inclusiveness.

As last experiment of this kind, the benchmark mixture proposed by Crouse in [60], which parameters are reported in (4.29), is reduced from $n^a = 10$ to $n^b = 4$ components; the results are reported in Fig. 5.3.

By setting $n^b = 4$, interesting choices are made by each of the algorithms, exhibiting more explicitly the corresponding features. The D_{RKL} again maps the 4 main peaks, exhibiting a pruning like behavior for the density region in the center. The D_{W2}, as mentioned, tries again to preserve the geometry of the mixture, this time, though, with a rather inclusive behavior. The D_{I2},

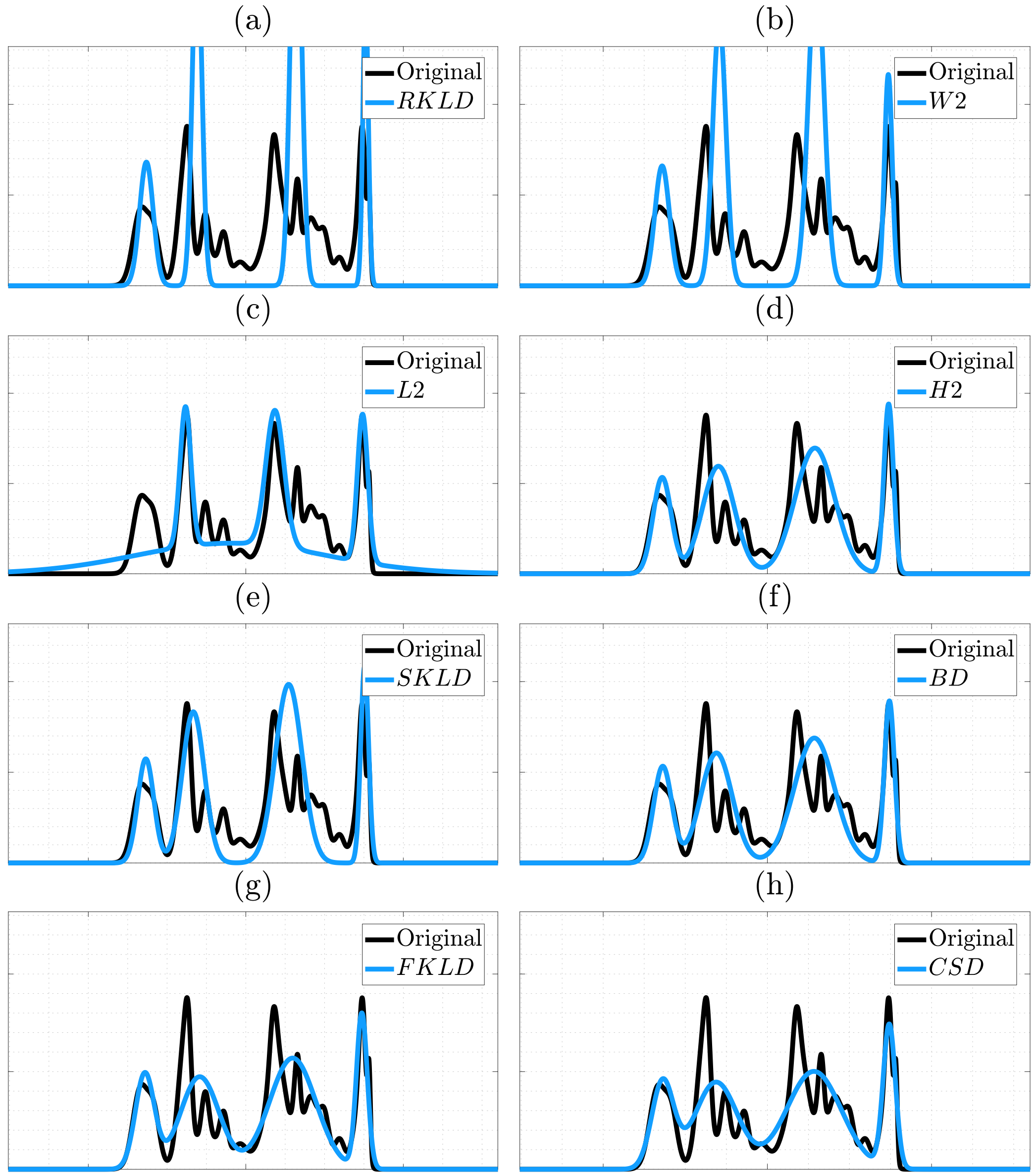
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)

Figure 5.2: GM of size $n^a = 40$ reduced to $n^b = 4$ by consistent C_D-based reduction for several D-measures. (a). $C_{D_{RKL}}$-based reduction. (b). $C_{D_{W2}}$-based reduction. (c). $C_{D_{L2}}$-based reduction. (d). $C_{D_{H2}}$-based reduction. (e). $C_{D_{SKL}}$-based reduction. (f). C_{D_B}-based reduction. (g). $C_{D_{FKL}}$-based reduction. (h). $C_{D_{CS}}$-based reduction.

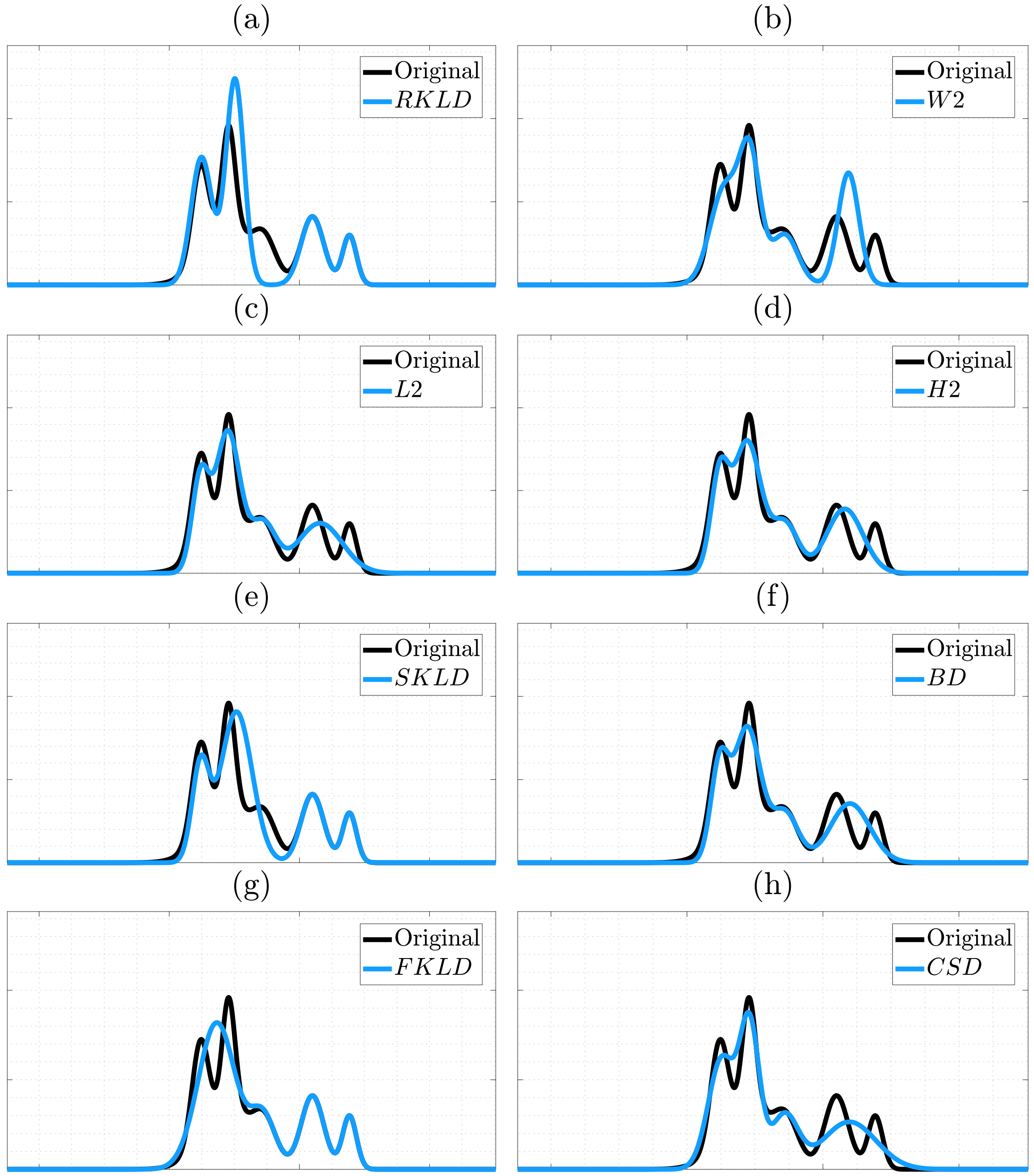

Figure 5.3: GM of size $n^a = 10$ reduced to $n^b = 4$ by consistent C_D-based reduction for several D-measures. (a). $C_{D_{RKL}}$-based reduction. (b). $C_{D_{W2}}$-based reduction. (c). $C_{D_{L2}}$-based reduction. (d). $C_{D_{H2}}$-based reduction. (e). $C_{D_{SKL}}$-based reduction. (f). C_{D_B}-based reduction. (g). $C_{D_{FKL}}$-based reduction. (h). $C_{D_{CS}}$-based reduction.

D_{H2}, D_{SKL}, D_B and D_{FKL} are coherent with the previous experiments. As announced, the D_{CS} can be puzzling. In Fig. 5.3 it favors the preservation of the main cluster of components, by instead spreading the covariance over the two isolated components. Here it is possible to observe closer the similarity with the D_B where, though, the covariances take larger magnitudes.

5.1.2 Consistent vs inconsistent approximations: a Monte Carlo study

To better figure out the numerical aspects of the consistency, in this section some Monte Carlo (MC) tests will be reported. Let us consider $N = 200$ 1-dimensional GMs, of size $n^a = 30$, generated out of $N_c = 6$ clusters as in 10, for parameters $\beta = 30$, $\delta_1 = 0.5$, $\delta_2 = 0.3$, and reduced to $n^b = 5$ components according to algorithm 4, for each D-measure listed in Table 5.1; in addition, a refinement is performed consistently for each algorithm as in Algorithm 8. Each reduction algorithm outcome is evaluated according to all of the corresponding induced C_D, and execution times are also reported. The average performances are evaluated over the $N = 200$ mixtures (Monte Carlo runs) and reported in Table 5.2. For each induced C_D (column), the value corresponding to the best performing algorithm (row) is reported in bold.

Monte Carlo average results for $N = 200$ 1-dimensional GMs									
Algo: C_D-red+ C_D-ref	Avg $C_{D_{RKL}}$	Avg $C_{D_{W2}}$	Avg $C_{D_{L2}}$	Avg $C_{D_{H2}}$	Avg $C_{D_{SKL}}$	Avg C_{D_B}	Avg $C_{D_{FKL}}$	Avg $C_{D_{CS}}$	Avg Time (s)
D_{RKL}	**2.0233**	2.5353	0.3555	0.2847	3.1307	0.5443	4.2360	1.0329	0.0683
D_{W2}	3.5484	**1.9587**	0.2906	0.2369	2.5393	0.4068	1.5302	0.7704	0.0834
D_{L2}	40.1494	55.9620	**0.1756**	0.2446	23.3020	1.0948	2.6907	1.9759	0.5480
D_{H2}	8.1315	2.7304	0.2060	**0.1870**	4.4763	0.2725	0.7933	0.4499	0.4472
D_{SKL}	2.5158	2.5430	0.2338	0.2185	**1.7808**	0.3132	1.0458	0.5680	0.1351
D_B	5.1966	2.7973	0.2055	0.1916	2.9369	**0.2413**	0.6719	0.3669	0.2611
D_{FKL}	7.9542	3.1132	0.2064	0.1964	4.3023	0.2478	**0.6373**	0.3546	**0.0591**
D_{CS}	10.6955	4.1269	0.2050	0.2079	5.6889	0.2610	0.6688	**0.3400**	0.1434

Table 5.2: Consistent full reduction pipelines applied on $N = 200$ 1-dimensional mixtures of size $n^a = 30$, and evaluated according to each induced average C_D and time.

By looking at Table 5.2, it is possible to observe how, according to each induced C_D measure, the corresponding algorithm achieves the best performance. Moreover, the smallest average time is employed by the D_{FKL}-induced algorithm.

Note: when discussing computational times of MR algorithms, the average time is not a reliable metric; since the employed time depends significantly on the implementation, the computational complexity should be the real metric considered in such kind of comparisons. For the above (and following) experiments, the same coding scheme has been employed for each of the algorithms.

The author wants to recall that many of the reported algorithms rely on FPI methods in order to evaluate the barycenters. If D-measures like the D_{I2} or the D_{H2} are considered, for which the D-barycenters may not exist unique, the corresponding FPI algorithm convergence can take many iterations to reach a sub-optimal solution. When the latter happens, there is no guarantee that the resulting approximation will achieve the best performance w.r.t. the relative induced C_D, especially if compared to other D-measures similar in the features, but more numerically robust. Another fact which should be briefly discussed, but which would require a chapter on its own, is how good of an approximation the induced C_D is in terms of the underlying, potentially intractable, D-measure. As discussed in section 4.4, the $C_{D_{FKL}}$ appears to be a good approximation for the D_{FKL} between mixtures, mostly for small deviations. This might not be true for other D-measures, but the corresponding discussion is beyond the goals of this dissertation. In any case, if an induced C_D is a rather coarse approximation for the base D-measure, there are no guarantees that the corresponding reduction algorithm will achieve the best performances in terms of the underlying dissimilarity. From this perspective, it may be reasonable to consider the surrogate dissimilarities C_D as measures on its own; of course, it would be desirable that the C_D represents a tight approximation for the base dissimilarity between mixtures.

Let us now consider another MC experiment where $N = 100$ 6-dimensional GMs of size $n^a = 30$, generated around $N_c = 6$ main clusters for parameters $\beta = 30$, $\delta_1 = 0.7$ and $\delta_2 = 0.5$, are reduced to $n^b = 5$ components as done in the previous experiment. The results are reported in Table 5.3; in this case though, since the magnitudes of each C_D can be very different, a suitable scaling factor will be used and reported in each column, when employed.

Monte Carlo average results for $N = 100$ 6-dimensional GMs									
Algo: C_D-red+ C_D-ref	Avg $C_{D_{RKL}}$ $(\cdot 10^{-2})$	Avg $C_{D_{W2}}$ $(\cdot 10^{-2})$	Avg $C_{D_{L2}}$ $(\cdot 10^{5})$	Avg $C_{D_{H2}}$	Avg $C_{D_{SKL}}$ $(\cdot 10^{-2})$	Avg C_{D_B}	Avg $C_{D_{FKL}}$ $(\cdot 10^{-1})$	Avg $C_{D_{CS}}$ $(\cdot 10^{-1})$	Avg Time (s)
D_{RKL}	**0.2266**	1.9985	1.8890	0.8959	0.2794	5.7026	3.3188	1.1131	0.1172
D_{W2}	0.2850	**1.6793**	0.8656	0.8613	0.2273	4.4265	1.6880	0.8621	1.1835
D_{L2}	3.9870	25.9594	**0.3729**	0.7802	2.0985	4.9016	1.1451	0.7750	0.7821
D_{H2}	0.6130	3.9147	0.5933	**0.6772**	0.4576	6.1183	2.5844	1.1711	2.3946
D_{SKL}	0.2609	1.9578	0.6504	0.8001	**0.1639**	2.2630	0.6695	0.4138	1.4502
D_B	0.5372	2.9291	0.5623	0.7271	0.2898	**1.4424**	0.3795	0.2058	0.5152
D_{FKL}	0.6430	3.5466	0.5591	0.7396	0.3446	1.4906	**0.3681**	0.2030	**0.0940**
D_{CS}	0.8857	4.9151	0.5574	0.7554	0.4702	1.5476	0.3830	**0.1918**	0.8456

Table 5.3: Consistent full reduction pipelines applied on $N = 100$ 6-dimensional mixtures of size $n^a = 30$, and evaluated according to each induced average C_D and time.

Again, by employing a consistent full reduction pipeline, each of the algorithms achieves the best performance according to the corresponding induced C_D. In terms of time, the $C_{D_{FKL}}$-based alternative confirms to be the most efficient one.

As last test, let us consider $N = 100$ 2-dimensional mixtures of size $n^a = 20$, generated according to $N_c = 6$ clusters for parameters $\beta = 15$, $\delta_1 = 0.4$, $\delta_2 = 0.3$, which are reduced to $n^b = 5$ components by means of the $C_{D_{FKL}}$-based greedy reduction, the Williams' algorithm (reported in section 3.5), and the $C_{D_{L2}}$-based greedy reduction algorithm. The refinements are omitted. As performance metrics, the three algorithms will be evaluated in terms of $C_{D_{FKL}}$, numerical D_{FKL}, D_{ISE}, which is recalled to possess a closed form in the mixture case, and the $C_{D_{L2}}$. Such an experiment involves two consistent hybrid algorithms, respectively the $C_{D_{FKL}}$ and the $C_{D_{L2}}$ based ones, and an inconsistent incrementally global algorithm, that is the Williams; the latter reduces incrementally the mixture by evaluating the D_{L2}, rather than the $C_{D_{L2}}$, between contiguous models, by either computing the D_{FKL}-barycenter between two components or by pruning the component introducing the least D_{L2}.

By looking at Table 5.4, it is possible to obtain several insights. First of

Monte Carlo average results for $N = 100$ 2-dimensional GMs					
Algorithm	Avg $C_{D_{FKL}}$	Avg D_{FKL}	Avg D_{ISE} ($\cdot 10^1$)	Avg $C_{D_{L2}}$ ($\cdot 10^1$)	Avg Time (s)
$C_{D_{FKL}}$-red	**1.1372**	**0.2871**	0.2422	0.6129	**0.0243**
Williams	6.8898	3.4222	0.2089	0.5614	6.1245
$C_{D_{L2}}$-red	5.2857	2.6808	**0.2013**	**0.4488**	0.6148

Table 5.4: Comparison between $C_{D_{FKL}}$-based, Williams' and $C_{D_{L2}}$-based reduction algorithms evaluated over several metrics.

all, the C_D-based reduction algorithms, as expected, achieve the best performances in terms of the corresponding induced C_D measures. In addition, such algorithms yield also very good performance w.r.t. the underlying corresponding D-measures between mixtures. The Williams' algorithm does not perform as good, for a significantly higher computational time, due to the following reasons:

- Inconsistency: such an algorithm optimizes the D_{ISE} in an incrementally global manner, but the merging actions are evaluated in terms of D_{FKL}-barycenters, hence optimizing, at the same time, the D_{FKL}.
- Pruning: allowing pruning may be a good alternative to merging if the chosen D-measure has an exclusive behavior. Nonetheless, in terms of D_{FKL}, this may lead to approximations preserving little to none density in regions where the original model has a non-zero mass: the corresponding D_{FKL} can even go to ∞. The D_{FKL} is a very inclusive D-measure, and does not evaluate well for approximations having a non-overlapping support with the original model.
- Merging: since merging is performed according to D_{FKL}, rather than D_{L2}, the overall approximation achieves, in general, inferior accuracy w.r.t. the D_{ISE} if compared with an equivalent algorithm where merging is done by means of D_{L2}-barycenter. It is important to recall that the D_{L2}-barycenter and the D_{L2}-BSDA do coincide, hence the D_{L2}-barycenter is the best merging action according to both the $C_{D_{L2}}$ and the D_{ISE} between mixtures.

Although the $C_{D_{L2}}$-based reduction optimizes an upper bound on the D_{ISE} rather than the D_{ISE} itself, the merging action is done consistently, hence the algorithm is expected to recover, in part, the accuracy due to its hybrid structure. In any case, the computational time is considerably inferior, even if particularly efficient implementations are used for the Williams' algorithm (not in this example). In conclusion, if the goal is to optimize the D_{L2}, while consuming less resources, it may be reasonable to employ the reported hybrid algorithm rather than the Williams'.

5.2 Extended Object Tracking: Adaptive Gamma Gaussian Inverse-Wishart Mixture Reduction

In this section will be investigated the problem of reducing a specific Random Finite Set (RFS)-based [8] uncertainty representation, often used in the context of extended object tracking, which falls under the name of gamma Gaussian inverse Wishart (GGIW) intensity. Extended Object Tracking (EOT) is a broad field, and discussing it in detail would require the introduction of several additional concepts; in this regard, the author decided to limit the argumentation to the problem of reducing a GGIW intensity [19], which represents another contribution provided by this dissertation. However, a comprehensive discussion about EOT can be found in the work of Koch [81].

5.2.1 Hypothesis management in extended object tracking

Multiple target tracking lies at the core of many important engineering achievements such as radar defence systems, air traffic surveillance systems and autonomous cars. The task of a multiple target tracker involves dealing with uncertainty in the number of targets, origin of the measurements, imperfections of the sensor and the target motion [11, 82]. As anticipated in Sec. 2.10, the uncertainty under investigation can be handled by considering a finite number of possible alternative events at each time step, called *hypotheses*. For example, in multiple target tracking, a single measurement can result from either a target for which a track has already been initialized or a new target or it can be a false alarm [13, 83]. In maneuvering target tracking the target motion can obey a mathematical model which can switch in a predefined set of stochastic difference/differential equations [4, Ch. 11].

When the uncertainty represented by different hypotheses cannot be resolved immediately, which is usually the case in a real life scenario, a Bayesian target tracker would yield probability density or intensity functions which are in the form of normalized or unnormalized mixtures of component densities defined over the state space. Every component of such a mixture/intensity would then correspond to an alternative hypothesis sequence in time. Unfortunately, as the time progresses, the number of such possible hypothesis sequences, hence the number of mixture components, would increase exponentially, which would make processing and storing such mixtures impossible. As a result one has to resort to MR algorithms, which have been deeply discussed in Chapters 3 and 4 of this dissertation.

In target tracking, the targets which can result in multiple measurements in a single sensor report are called *extended targets*. The last one and a half decades have seen a plethora of recent advances in both the theory and practice of extended target tracking (ETT) [84]. A recent mixture/intensity type that has gained popularity in extended target tracking literature is the Gamma Gaussian inverse-Wishart (GGIW) mixture/intensity [19,20,24,26]. In a GGIW mixture/intensity used in an Extended Target Tracking (ETT) context

- the components would represent the estimated statistics of the extended target(s);

- the weights would represent either the hypothesis probabilities or expected number of extended targets (belonging to a specific component) depending on whether the corresponding mixture/intensity is normalized (mixture case) or unnormalized (intensity case);

- the gamma part of the components would hold the statistics for the number of measurements generated by the corresponding extended target(s);

- the Gaussian part of the components would hold the statistics for the kinematics states of the extended target(s);

- the inverse Wishart part of the components would hold the extent statistics of the corresponding extended target(s).

Greedy algorithms to reduce Gamma mixtures and Gaussian inverse Wishart mixtures were given in [27] and [85], respectively, where symmetric Kullback

Leibler divergence is used as the merging criterion and Kullback Leibler divergence is minimized to find the merged component. In [19], D_{FKL}-consistent reduction and refinement algorithms have been provided for the GGIW case.

For the sake of discussion, the theory regarding GGIW reduction is reported in the first part of this section; in addition, adaptive reduction algorithms will be proposed in order to provide the automatic selection of the number of GGIW hypotheses.

5.2.2 GGIW representations

Let us consider a weighted sum of n densities of the form:

$$\begin{aligned} p(\xi|\Theta) \triangleq \tilde{\boldsymbol{w}}^T \boldsymbol{\zeta}(\chi, x, \mathcal{Y}|\boldsymbol{\theta}) &= \sum_{i=1}^{n} \tilde{w}_i \zeta_i(\xi|\theta_i) \\ &= \sum_{i=1}^{n} \tilde{w}_i \gamma(\chi|\kappa_i, \omega_i) \nu(x|\mu_i, \Sigma_i) \varphi(\mathcal{Y}|V_i, v_i), \end{aligned} \tag{5.1}$$

where $\xi = (\chi, x, \mathcal{Y}) \in \mathbb{R}_+ \times \mathbb{R}^d \times S^d_{++}$, γ_i is the i-th gamma density as defined in (2.57), ν_i is the i-th Gaussian density as defined in (2.55) and φ_i is the i-th inverse Wishart density as defined in (2.60). For a set of GGIW densities, the corresponding parameter set will be $\boldsymbol{\theta} = \{\boldsymbol{\kappa}, \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{V}, \boldsymbol{v}\} \in \mathcal{H}^\theta_n = \mathbb{R}^n_+ \times \mathbb{R}^n_+ \times (\mathbb{R}^d)^n \times (S^d_{++})^n \times (S^d_{++})^n \times \mathbb{R}^n_+$. $\tilde{\boldsymbol{w}} = [\tilde{w}_1, ..., \tilde{w}_n]^T \in \mathbb{R}^n_+$ will be used for the intensity case, that is $p(\xi|\Theta) \in \mathcal{Q}_{\text{int}}$, whereas $\boldsymbol{w} \in \Delta^{n-1}$ will be used for the mixture case, that is $p(\xi|\Theta) \in \mathcal{Q}_{\text{mix}}$. ζ_i is the i-th GGIW density and $\boldsymbol{\zeta}(\xi|\boldsymbol{\theta}) = [\zeta_1(\xi|\theta_1), ..., \zeta_n(\xi|\theta_n)]^T$. Hypotheses as the GGIWs are called *product* densities, since they are obtained as the product of several simpler densities. For the sake of discussion, terms as *joint* or *overall* will be used to refer to the product density as a whole, while terms like *marginal* will be used to refer to the single densities involved in the product.

5.2.3 $C_{D_{FKL}}$-based reduction

The CTD-based reduction framework represents a totally general approach which can be used to deal both with mixtures and intensities; the only requirements are that one can compute the pairwise dissimilarities between hypotheses, and can evaluate the D-barycenter of a set of components. At the end of the previous chapter, it has been discussed the fact that the D_{FKL}

possesses many useful properties which make it a core measure in the MRP when addressed by means of OTT. In this regard, it is worth to list another interesting property which simplifies considerably the GGIW reduction problem.

Separability property of the D_{FKL}

In the case of product densities, the D_{FKL} possesses a useful property, namely the *separability*, also known as *additivity for product distributions* [79]; given two GGIW densities $\zeta_i = \gamma_i(\chi)\nu_i(x)\varphi_i(\mathcal{Y})$ and $\zeta_j = \gamma_j(\chi)\nu_j(x)\varphi_j(\mathcal{Y})$, the following relation holds:

$$D_{FKL}(\zeta_i \| \zeta_j) = D_{FKL}(\gamma_i \| \gamma_j) + D_{FKL}(\nu_i \| \nu_j) + D_{FKL}(\varphi_i \| \varphi_j). \tag{5.2}$$

The proof is reported in Appendix B.2.6. The three terms in (5.2) have been defined respectively in (3.13), (3.12) and (3.14).

D_{FKL}-barycenter of GGIW densities

Let us consider a set of n GGIW densities $(\tilde{\boldsymbol{w}}, \boldsymbol{\zeta}) = \{\tilde{w}_i, \zeta_i\}_{i=1}^n$; by exploiting property (5.2), it is trivial to obtain that:

$$\hat{\zeta}(\xi) = \hat{\gamma}(\chi)\hat{\nu}(x)\hat{\varphi}(\mathcal{Y}), \tag{5.3}$$

where $\hat{\gamma}$ is the D_{FKL}-barycenter of the gamma marginal as defined in (3.86), $\hat{\nu}$ is the D_{FKL}-barycenter of the Gaussian marginal as defined in (3.81) and $\hat{\varphi}$ is the D_{FKL}-barycenter of the inverse Wishart marginal as defined in (3.82); in other words, the D_{FKL}-barycenter of a set of GGIW densities can be factorized as the product of the corresponding gamma, Gaussian and inverse Wishart marginal barycenters of the given intensity.

GGIW reduction and refinement

The separability property allows to compute efficiently both the pairwise D_{FKL} between GGIW hypotheses and the D_{FKL}-barycenters of a set of GGIW components; thus said, one can resort to the theory presented in Chapter 4 in order to reduce/refine a mixture/intensity of GGIW densities. Nonetheless, when dealing with intensities, there are few facts which should be kept in mind; first of all, it is important to remark that the D_{FKL} is a *statistical divergence*, that is a way to measure how distant two probability distributions are; this implies that such a measure should be considered, in its standard definition, for mixture densities rather than intensities. If the latter are considered, it would not be obvious how the corresponding D_{FKL} should be computed. For practical applications, though, the optimal transport framework allows to perform reduction and refinement of intensities without the need to normalize the corresponding weights, given that the cost matrix is built with dissimilarities between pairs of densities (the weights plays no role in the cost matrix computation).

If compared to the marginal case, the GGIW mixture/intensity reduction problem is, in general, more difficult: performing a reduction step according to a cost-based criterion, might lead to situations where the cost for one of the marginal densities is optimal, whereas it results to be sub-optimal for the other two. This yields in general overall good approximations, but in the marginal perspective potentially inferior solutions often occur. In this regard, the availability of accurate reduction algorithms is particularly important when dealing with product hypotheses.

5.2.4 Adaptive reduction of GGIW hypotheses

A contribution of this dissertation has been the GGIW reduction for a fixed number of reduced mixtures components, reported in the work [19]; nonetheless, given the availability of adaptive reduction theory 4.4, already discussed broadly in Sec. 4.4, the author decided to report a discussion about adaptive GGIW intensity reduction rather than the fixed number of hypotheses case. In this regard, all the necessary fundamentals will be reported following.

Proposition 5.2.1 (Hybrid bound and CTD equality for GGIW intensities)**.** *Let us consider a GGIW intensity* $p^a = (\tilde{\boldsymbol{w}}^a)^T \boldsymbol{\zeta}^a$ *and its reduced-by-one model* $\tilde{p}^a_{i,j}$ *obtained as in* (4.16)*. If* $\tilde{p}^a_{i,j}$ *is obtained by merging the pairs of components* (i, j) *of* p^a *for which the bound* $B_D(\tilde{w}^a_i \zeta^a_i, \tilde{w}^a_j \zeta^a_j)$ (4.17) *is minimum, then theorem 4.3.1 holds, that is:*

$$C_D(p^a \| \tilde{p}^a_{i,j}) = B_D(\tilde{w}^a_i \zeta^a_i, \tilde{w}^a_j \zeta^a_j). \tag{5.4}$$

Moreover, if D is a separable measure, it holds:

$$\begin{aligned} C_D(p^a \| \tilde{p}^a_{i,j}) &= B_D(\tilde{w}^a_i \zeta^a_i \| \tilde{w}^a_j \zeta^a_j) = \\ &= B_D(\tilde{w}^a_i \gamma^a_i \| \tilde{w}^a_j \gamma^a_j) + B_D(\tilde{w}^a_i \nu^a_i \| \tilde{w}^a_j \nu^a_j) + B_D(\tilde{w}^a_i \varphi^a_i \| \tilde{w}^a_j \varphi^a_j), \end{aligned} \tag{5.5}$$

that is the bound B_D is separable and the C_D between the mixture before and the one after a minimum cost merging action can be computed as the sum of the three marginal bounds. In addition, as done for the Gaussian case, the cumulative sum of the overall optimal bounds can be used as an estimate of the CTD between the original mixture/intensity and a corresponding reduced instance.

The proof of the bound separability for the GGIW case is reported in Appendix B.2.6.

Proposition 5.2.2 (Joint D_{FKL}-barycenter cost)**.** *Let us consider a GGIW intensity $p^a = (\tilde{\boldsymbol{w}}^a)^T \boldsymbol{\zeta}^a$ and its D_{FKL}-barycenter $\hat{\zeta}^a$ computed as in* (5.3)*. It is trivial to prove that:*

$$c(\hat{\zeta}^a|p^a) = C_{D_{FKL}}(p^a\|\hat{\zeta}^a) = m_{D_{FKL}}(\hat{\zeta}^a|\tilde{\boldsymbol{w}}^a, \boldsymbol{\zeta}^a) = \sum_{i=1}^{n^a} \tilde{w}_i^a D_{FKL}(\zeta_i^a\|\hat{\zeta}^a). \quad (5.6)$$

By exploiting the separability property (5.2)*, one obtains the following relations:*

$$\begin{aligned} c(\hat{\zeta}^a|p^a) = C_{D_{FKL}}(p^a\|\hat{\zeta}^a) &= \sum_{i=1}^{n^a} \tilde{w}_i^a D_{FKL}(\zeta_i^a\|\hat{\zeta}^a) = \\ &= \sum_{i=1}^{n^a} \tilde{w}_i^a D_{FKL}(\gamma_i^a\|\hat{\gamma}^a) + \sum_{i=1}^{n^a} \tilde{w}_i^a D_{FKL}(\nu_i^a\|\hat{\nu}^a) + \sum_{i=1}^{n^a} \tilde{w}_i^a D_{FKL}(\varphi_i^a\|\hat{\varphi}^a) = \\ &= C_{D_{FKL}}((\tilde{\boldsymbol{w}}^a)^T\boldsymbol{\gamma}^a\|\hat{\gamma}^a) + C_{D_{FKL}}((\tilde{\boldsymbol{w}}^a)^T\boldsymbol{\nu}^a\|\hat{\nu}^a) + C_{D_{FKL}}((\tilde{\boldsymbol{w}}^a)^T\boldsymbol{\varphi}^a\|\hat{\varphi}^a) = \\ &= c(\hat{\gamma}^a|(\tilde{\boldsymbol{w}}^a)^T\boldsymbol{\gamma}^a) + c(\hat{\nu}^a|(\tilde{\boldsymbol{w}}^a)^T\boldsymbol{\nu}^a) + c(\hat{\varphi}^a|(\tilde{\boldsymbol{w}}^a)^T\boldsymbol{\varphi}^a). \end{aligned} \quad (5.7)$$

that is the overall D_{FKL}-barycenter cost can be obtained as the sum of the three marginal D_{FKL}-barycenter costs.

The previous propositions can be exploited to apply the adaptive reduction theory, presented in Sec. 4.4, to the case of GGIW intensities. For product densities, there are several ways to perform an adaptive reduction; the first one is to consider the application of the two criteria (ATC and SMC) presented in Sec. 4.4 directly on the corresponding joint curves. In alternative, one could consider different thresholds for each of the marginal curves. Nonetheless, it is important to stress out the following fact: what represents an optimal choice from an overall perspective, may represent a sub-optimal step for the marginal case.

For the first part of this section, the discussion will be focused on the ATC case, since few results were already available when the SMC was formulated, and to unclutter plots involving many quantities. Given the optimal joint bound values, the cumulative cost built accordingly is a cumulative RTL curve, since equation (4.22) holds. Nonetheless, the cumulative curves built on the marginal bounds, corresponding to the optimal overall bound, do not represent true cumulative RTL curves. This is due to the fact that, marginally, the choice may not be optimal, hence the strict inequality holds in equation (4.24). In any case, a normalization of the marginal curves can still be obtained by considering the marginal D_{FKL}-barycenter costs, that is one can define the following marginal curves:

$$s_{B,\gamma}^{(m)} = \frac{\sum_{l=n^a}^{m} \breve{B}_{D_{FKL},\gamma}^{(l)}}{c(\hat{\gamma}^a | (\tilde{\boldsymbol{w}}^a)^T \boldsymbol{\gamma}^a)}, \tag{5.8}$$

$$s_{B,\nu}^{(m)} = \frac{\sum_{l=n^a}^{m} \breve{B}_{D_{FKL},\nu}^{(l)}}{c(\hat{\nu}^a | (\tilde{\boldsymbol{w}}^a)^T \boldsymbol{\nu}^a)}, \tag{5.9}$$

$$s_{B,\varphi}^{(m)} = \frac{\sum_{l=n^a}^{m} \breve{B}_{D_{FKL},\varphi}^{(l)}}{c(\hat{\varphi}^a | (\tilde{\boldsymbol{w}}^a)^T \boldsymbol{\varphi}^a)}. \tag{5.10}$$

$\breve{B}_{D_{FKL},\gamma}^{(l)}$, $\breve{B}_{D_{FKL},\nu}^{(l)}$ and $\breve{B}_{D_{FKL},\varphi}^{(l)}$ are respectively the bounds associated to the gamma, Gaussian and inverse Wishart marginals, underlying the optimal (overall) merging cost $\breve{B}_{D_{FKL}}$ corresponding to the l-th reduced order model; in other words, one can compute marginal cumulative curves as the sum of the marginal bounds $\breve{B}_{D_{FKL},(\cdot)}$ and then by normalizing over the corresponding (marginal) D_{FKL}-barycenter cost. Such curves, though, will not be as "smooth" as the overall (optimal) case, since, as mentioned, the minimum cost merging (done in a *joint* perspective) will not take into account the marginal optimality, hence sequential reduction steps could yield significantly inferior results for the single density. To better understand this fact, in Fig. 5.4 is reported a GGIW intensity reduction, where a threshold $\lambda_{\tilde{\mathcal{L}}} = 15\%$ has been used on the overall cumulative RTL curve to halt the reduction.

The three (normalized) marginal cumulative curves have also been plotted in order to show how sub-optimal merging affects the corresponding cumulative costs. Since for sub-optimal choices the strict inequality of (4.24) holds, adaptive reductions done according to marginal cumulative curves may re-

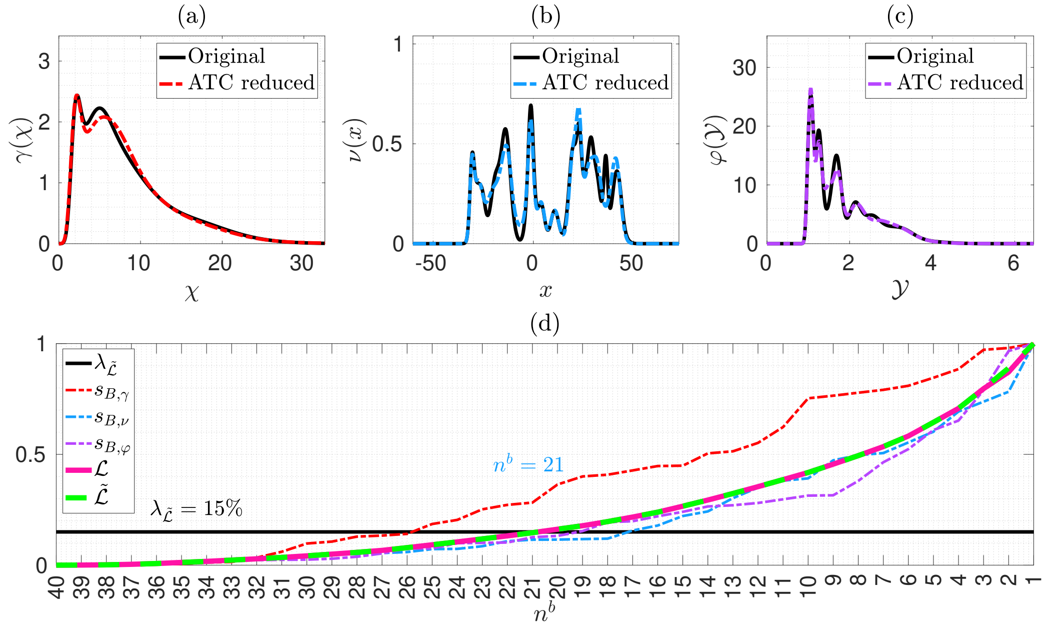

Figure 5.4: GGIW intensity of size $n^a = 40$ reduced adaptively by means of ATC for $\lambda_{\tilde{\mathcal{L}}} = 15\%$, corresponding to a reduced model of size $n^b = 21$. (a). gamma intensity reduction. (b). Gaussian intensity reduction. (c). inverse Wishart intensity reduction. (d). Joint RTL curves (dashed green and magenta), marginal cumulative curves respectively in dashed red, dashed light-blue, dashed purple.

sult to be particularly conservative, since significant values can be reached even for very low compression ratios. In this regard, let us assume that the reduction is halted if either one of the three marginal curves exceeds the accuracy threshold rather than the overall one. By looking at Fig. 5.4, it is possible to observe that the gamma marginal curve (in red) exceeds the accuracy threshold $\lambda_{\tilde{\mathcal{L}}}$ at around $n^b = 26$ components, instead of $n^b = 21$ as in the overall case. The global joint RTL curve has been plotted again since something mentioned happened; for the reduction step from $n^b = 3$ to $n^b = 2$, the $\tilde{\mathcal{L}}$ detached from $\mathcal{L}$ (meeting again in $n^b = 1$). This phenomena is rather rare in single marginal cases, but can occur more frequently when product distributions are considered. In any case, as already discussed, such detachment happens for very low model orders which, in general, are never explored from the adaptive reduction algorithm.
Let us now consider another example where the ATC is considered for $\lambda_{\tilde{\mathcal{L}}} = 20\%$, reported in Fig. 5.5.

As it can be observed, the Gaussian cumulative curve (dashed light blue) majorizes slightly the overall RTL curve for low values of reduction ratios; from around $n^b = 27$, such a trend occurs in the inverse Wishart marginal. Regarding the marginal case for the SMC, at this stage it has not beed investigated yet, but the author believes that it might not be worth to be used: since the SMC is more sensible to symmetry/modality modifications (which now can happen over three marginals), it would be reasonable to expect even more conservative reductions if compared to the marginal RTL case.

For the remainder of this discussion, only the overall perspective will be considered in the GGIW intensity reduction problem; nonetheless, if one wants to perform a reduction where different thresholds on the marginals are provided, it still can be done by considering the cumulative curves (5.8) (in the ATC case, or the equivalents for the SMC case): such a choice specifically depends on the application of interest.

GGIW intensity reduction by ATC and SMC

To further investigate the GGIW intensity reduction problem, let us now consider both the presented halting criteria in 4.4; again, for the Gaussian case, the components are generated according to $N_c = 6$ clusters rather than with uniform probability over the space.

From several other experiments, and as briefly mentioned, it seems that

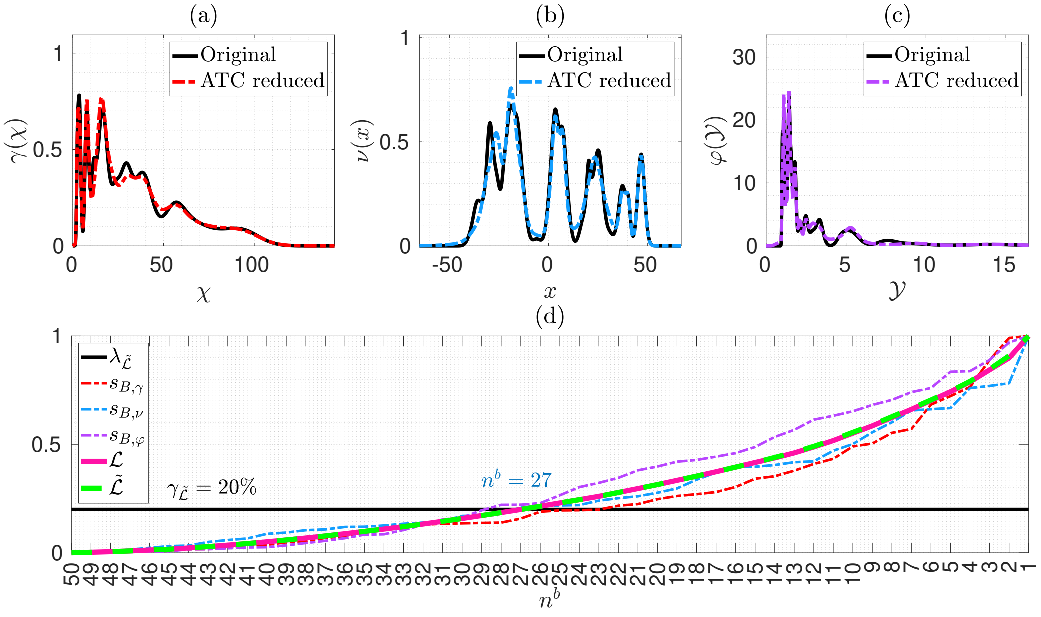

Figure 5.5: GGIW intensity of size $n^a = 50$ reduced adaptively by means of ATC for $\lambda_{\tilde{\mathcal{L}}} = 20\%$, corresponding to a reduced model of size $n^b = 27$. (a). gamma intensity reduction. (b). Gaussian intensity reduction. (c). inverse Wishart intensity reduction. (d). Joint RTL curves (dashed green and magenta), marginal cumulative curves respectively in dashed red, dashed light-blue, dashed purple.

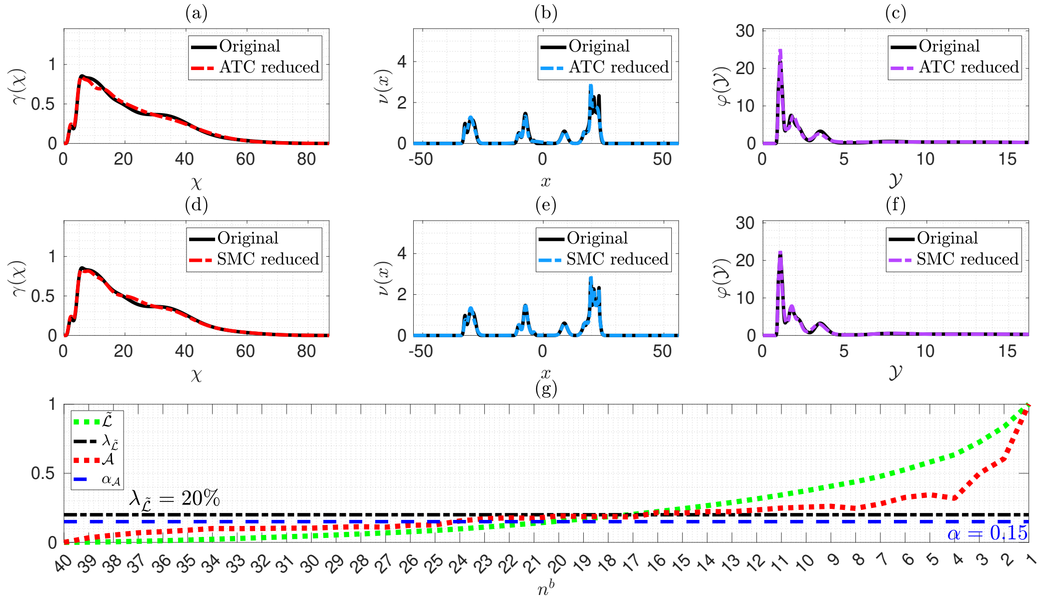

Figure 5.6: 40-component GGIW intensity reduced by means of ATC, for $\lambda_{\tilde{\mathcal{L}}} = 20\%$, yielding $n^b = 18$, and by means of SMC, for $\alpha_{\mathcal{A}} = 0.15$, yielding $n^b = 25$. (a)-(d). gamma intensity reduction. (b)-(e). Gaussian intensity reduction. (c)-(f). inverse Wishart intensity reduction. (g). Halting criteria.

the SMC criterion is rather sensible in the product densities case: marginal merging actions may not be optimal, hence the quantity (4.28) can increase rapidly. In this regard, it might be worth to consider a value of $\alpha_{\mathcal{A}} = 0.20$.

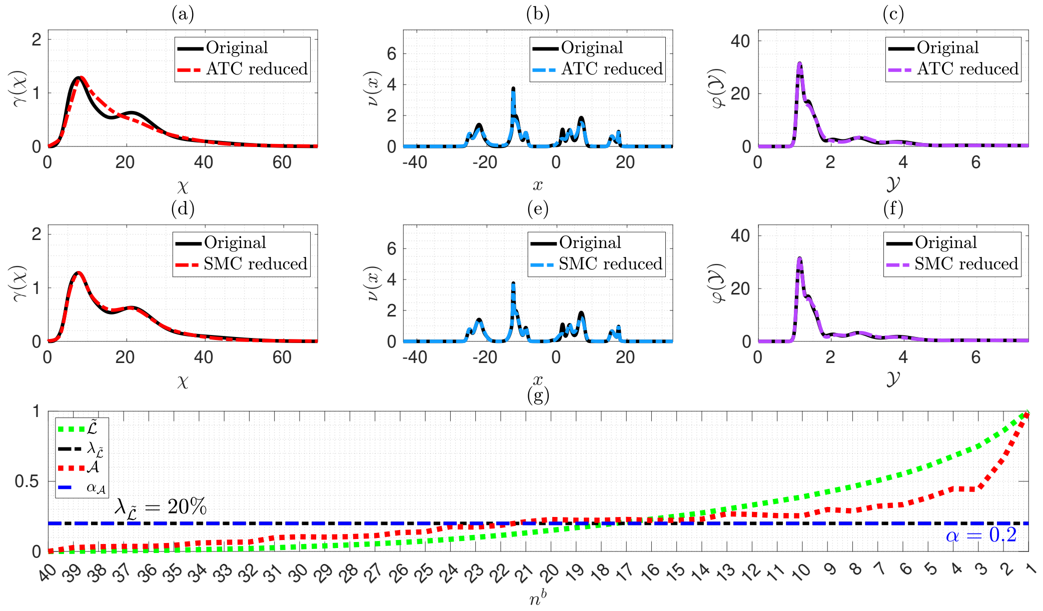

Figure 5.7: 40-component GGIW intensity reduced by means of ATC, for $\lambda_{\tilde{\mathcal{L}}} = 20\%$, yielding $n^b = 18$, and by means of SMC, for $\alpha_{\mathcal{A}} = 0.2$, yielding $n^b = 22$. (a)-(d). gamma intensity reduction. (b)-(e). Gaussian intensity reduction. (c)-(f). inverse Wishart intensity reduction. (g). Halting criteria.

With the given thresholds, the two algorithms seem to perform comparably; nonetheless, a key concept which is worth to be recalled, is the fact that the performance of an algorithm should not be evaluated solely on the single reduction realization, but it should be either considered a MonteCarlo test or real world data on which a tuning phase has been done. In any case, single synthetic realizations can help to figure out the general algorithm behavior. Even in the GGIW case, though, the SMC seems to exhibit an overall better performance w.r.t. the ATC.

Some notes on adaptive reduction and object tracking

In target tracking algorithms based on RFS, the sum of the intensity weights serves as estimate of the expected number of objects present in the field of view; many tracking algorithms based on the concept of Probability Hypothesis Density (PHD), resort often to threshold-based pruning in order to remove the unlikely state hypotheses. Nonetheless, significantly probable hypotheses could be redundant if originated by the same object, hence it would be reasonable to consider merging to further simplify the uncertainty description. In this regard, resorting to an adaptive reduction as discussed in this section, might help to reduce the number of hypotheses (model complexity) while preserving the expected number of objects. The author wants to remark that pruning is a destructive practice, since information is lost when hypotheses are removed. In contrast, merging-only reduction algorithms preserve all the information, which, at a first glance, might seem to be dangerous from a target tracking in presence of clutter perspective. Nonetheless, given the D_{FKL} features, the merging of very unlikely isolated hypotheses would generate new components spread in the covariance, which leads to an even lower corresponding importance in the next filter update (see Sec. 2.10). In this regard may be reasonable to at first perform an adaptive merging-only reduction, and then to perform a pruning where the weights are normalized over the determinant of the covariance matrix (Gaussian case), that is to consider:

$$\bar{w}_i = \frac{\tilde{w}_i}{\sqrt{|\Sigma_i|}}, \; i = 1, ..., n. \tag{5.11}$$

Such a normalization can help to remove both low weighted and covariance-spread components (the latter can result from the merging-only reduction algorithm). If, instead, the unlikely components appear near significant clusters, merging those is expected to barely influence the overall cluster modality. Given the features of the so far presented reduction algorithms, the author is confident that adaptive reduction of PHDs could improve the overall performance of tracking algorithms.

Preserving the inverse Wishart mixture mean

To conclude this section, it might be worth to address even the problem of reducing a mixture/intensity of inverse Wishart densities according to the D_{FKL} measure. The D_{FKL}-barycenter does not preserve mean value of an inverse Wishart mixture, in contrast with the gamma and Gaussian cases. Performing aggressive reductions may distort significantly the mean value of an inverse Wishart mixture/intensity; a possible mitigation of this problem is provided by the adaptive algorithms here presented since, by preserving the mixture/intensity modality, one could expect to preserve the corresponding statistics. Several tests, not reported here since not a core argumentation of this work, shown that both the ATC and SMC are capable of avoiding that the mean value of an inverse Wishart mixture is distorted significantly in a reduction problem.

5.3 Model Selection for Sample Data: Coupling Data Clustering with Mixture Reduction

Another application of the discussed adaptive reduction algorithms is the one of finding a suitable model order for sample data.

As discussed in Sec. 2.8, one can employ mixture models in order to characterize a given dataset: clustering algorithms as the EM (see 2.8) represent a possible way to fit mixture models over the data. Nonetheless, it is usually unclear how many components would be required to find a suitable description, and the evaluation of several model orders may require a lot of time; in addition, the EM algorithm is particularly sensible to the initialization. In Sec. 2.7 the problem of model selection has been discussed, and two information criteria have been reported, namely the AIC and the BIC. Both those criteria though, require all the model orders to be evaluated in order to find the best trade-off between model complexity and representation accuracy (given in terms of the likelihood function). For these reasons, the problem of fitting a mixture model to the data is not of easy solution.

In this section, the author proposes a very recent idea on the exploitation of adaptive reduction to find a suitable model order efficiently.

5.3.1 Coupling EM and adaptive reduction theory

The proposed idea relies on the coupling of any sample data clustering algorithm and an adaptive reduction. The EM is considered since it yields directly a mixture model and approximately provides the MLE of the mixture model over the data; in any case, any other algorithm can be considered, with the only requirement of yielding a mixture model after its execution is completed (e.g. K-means with cluster sample covariance computation). For the goals of this argumentation, only Gaussian mixtures will be considered, but the approach remains general.

Overfitting the data on purpose: is it a good idea?

Let us consider the problem of finding a probabilistic model for an unknown process of interest, from which observations can be obtained. Hence, given a set of corresponding samples, one way to provide a mathematical description could be to consider the Kernel Density Estimation (KDE)[3] approach [86, 87]; nonetheless, such a method relies on the choice of the kernel bandwidth, and the final outcome will be strongly influenced by such a choice. If a Gaussian kernel is chosen, the bandwidth is represented by the covariance matrix, and the overall distribution can be seen as a Gaussian mixture with as many components as the number of observations. Nevertheless, unless singular covariances are considered, it is not an easy task to find a suitable kernel bandwidth to approximate the distribution of interest. Moreover, a reduction performed on such "mixture" will also strongly depend on the chosen bandwidth. An alternative, as discussed, is to consider the EM algorithm, which is recalled to be a soft-clustering algorithm capable of fitting mixture models over the observed data. In this case, though, the problem of finding a suitable initialization, jointly with the number of components, can make the problem of fitting a model over the data particularly difficult, especially in high dimensions. Of course, if one has to find a model which best explains the given data, it would make more sense to overfit the data rather than underfitting it, but, from the *bias-variance dilemma* [3], one knows that a high model variance causes in general a loss of generalization ability; on the other hand, underfitting the data may lead to inferior prediction accuracy. Given a set of d-dimensional observations $\boldsymbol{x} = \{x_1, ..., x_N\}$, $x_i \in \mathbb{R}^d$, the adaptive reduction theory presented in Sec. 4.4 could be exploited as follows:

1. Select a high number of components and initialize randomly the EM algorithm (on-purpose overfitting).

2. Perform an adaptive reduction of the EM outcome as shown in Sec. 4.4.

[3] KDE is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the real unknown distribution are made, based on a finite data sample.

3. Refine the reduced order model by another instance of the EM algorithm.

With this approach, though, there are few problems; first of all, if too many initial components are provided, the EM could incur in singularities (e.g. a given component collapses onto a single observation, yielding a singular Gaussian density of zero covariance). Nonetheless, in the literature are present many workarounds to overcome this kind of situations, like resetting the singular component(s) somewhere else and keep iterating until convergence. Another problem, of conceptual nature, arises from the fact that the EM algorithm is a soft clustering approach, while the adaptive reduction theory here presented relies on hard-clustering theory. This may represent a problem since, during the descent, the adaptive algorithm will "detach from the data" and the subsequent reduction steps will be done in a hard-clustering perspective, hence deviating from what an equivalent soft-clustering algorithm would yield. For this reason, a final instance of EM is mandatory to both "re-attach the model to the data" and to provide a corresponding soft-clustering estimate.

In order to provide some insights on the performances of such an approach, let us consider a dataset containing $N = 3000$ observations generated according to $N_c = 20$ Gaussian components; the goal is to retrieve a final model for the data with a suitable order. To further discuss the features of this approach, the BIC will be evaluated for each of the reduced order model provided by the adaptive reduction. In Fig. 5.8 are reported the corresponding results.

As it can be observed, $n^a = 50$ components represent a significant unnecessary number, hence causing an overfitting of the data; nonetheless, the data modality seems to be preserved. By then using the SMC criterion with $\alpha_{\mathcal{A}} = 0.15$, the reduction stops for $n^b = 20$, which represents the desired number of components. On the other hand, the BIC curve suggests that the optimal value may be $n^b = 23$, a different value than the expected one: there is a specific motivation for such a behavior. Once the initialization is provided, that is the data has been encoded with a model, the adaptive reduction is performed until a halting occurs. Moreover, during the descent, the algorithm is detaching from data and working in a sort of simplified domain made of Gaussian components rather than samples. In such domain, though, the choices are chosen according to a hard-clustering criterion, and subsequent merging actions may yield solutions which differ significantly from a model

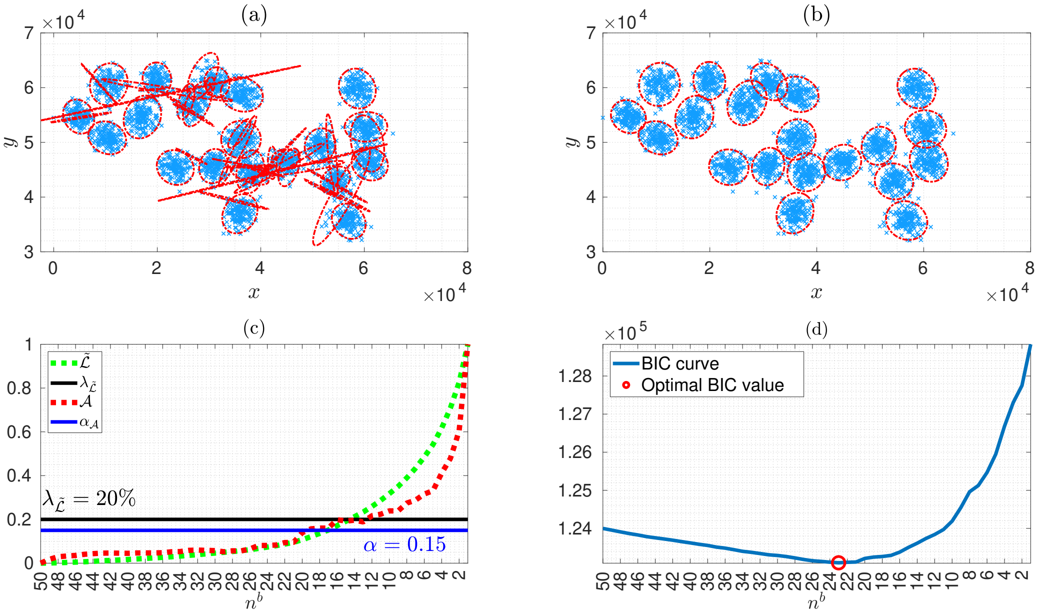

Figure 5.8: (a). Set of $N = 3000$ samples fitted by means of EM algorithm with $n^a = 50$ Gaussian components. (b). The $n^a = 50$ components are reduced adaptively and another pass of the EM is performed. (c). Halting criteria. (d). BIC values as function of model orders.

provided by a soft-clustering equivalent; in fact, the adaptively reduced model will not represent a good description of the data as the one would obtain by fitting a same order model with a good run of the EM; luckily, the final instance of EM usually fixes the introduced distortion, by yielding, in general, really accurate models with suitable orders. Thus said, the BIC reported in Fig. 5.8.(d) may be providing the wrong optimal order due to the fact that the corresponding evaluation is done over models which do not really explain well the data. A more rigorous usage of the BIC is by "re-attaching the model to the data" after each merging action; however, this would break the adaptive reduction theory, since altering the mixture from one reduction step to another would cause the ATC/SMC curves to miss the summation to one, hence making pointless to consider thresholds in the interval $[0, 1]$. In this regard, though, let us consider the same problem as in Fig. 5.8, but where the BIC is computed after performing an EM refinement "external" to the reduction, in the sense that the reduced order models involved in the descent are not modified, but the corresponding BIC is computed after an EM refinement of each model order. In Fig. 5.9 is reported the result.

As it can be observed, now the BIC correctly estimates the optimal number of components; in fact, the BIC should be evaluated for the likelihood-maximized model over the data, and not for a model yielded by the adaptive reduction. In addition, if one looks at both the corresponding BIC plots of Figures 5.8 and 5.9, something more can be evinced: the D_{FKL}-consistent greedy reduction seems to cause a smooth decrease of the BIC down to the optimal number. The logic behind this may rely on the following considerations: first of all, hierarchical models like reduced mixtures are, by construction, simpler models in terms of parameters which aim to preserve the original accuracy; in this regard, merging very similar components does not alter, in general, the mixture modality in a significant manner, hence allowing for a reduction of the parameter number with a small loss of information. If an overfitting process as the one proposed at the beginning of this section takes place, then it is reasonable to find several components which, in a greedy reduction, perspective, can be simplified without losing significant representation capabilities. Given that the most similar components are merged at each step according in a reduction like the one proposed in Sec. 6, one could expect a monotonically decreasing accuracy (smaller values of the likelihood function over the data), which, in the BIC case, is balanced by the number of parameters in the mixture, decreasing at each reduction step.

Nevertheless, the reported plots may be lucky realizations of random ex-

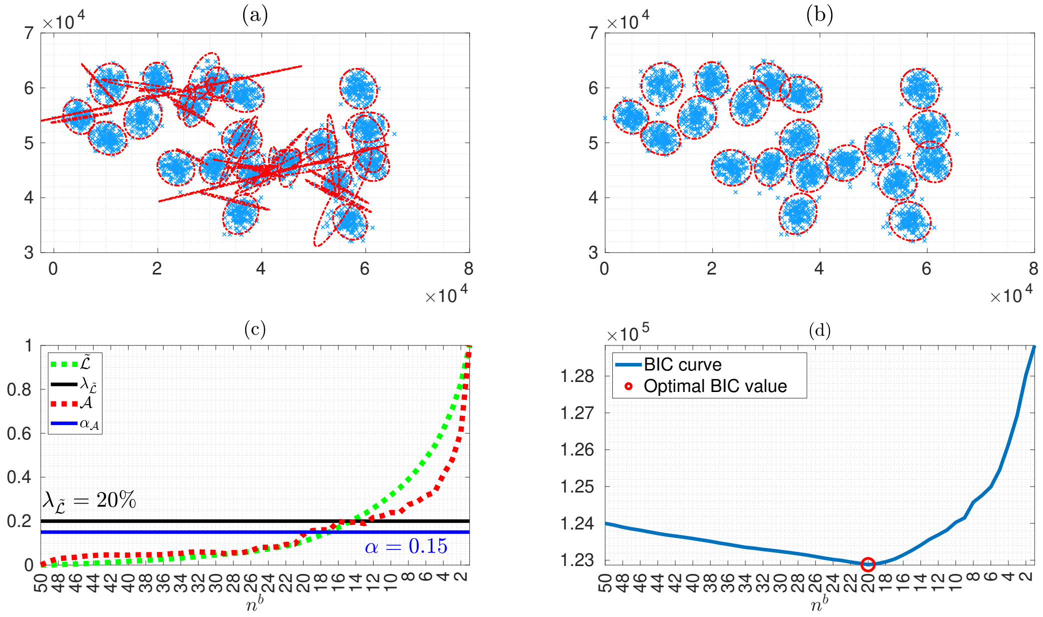

Figure 5.9: (a). Set of $N = 3000$ samples fitted by means of EM algorithm with $n^a = 50$ Gaussian components. (b). The $n^a = 50$ components are reduced adaptively and another pass of the EM is performed. (c). Halting criteria. (d). BIC values as function of model orders.

periments, since there is another fact regarding the BIC which should be remarked. Information criteria like the BIC tend to penalize the number of parameters in order to achieve an *Occam's razor* trade-off between accuracy and complexity. Nonetheless, by looking at the corresponding formula (2.103), it can be noticed that the penalization term depends on the number of observations (something that the AIC does not possess, hence providing in general more complex models). In high dimensions it may happen that the likelihood part totally obscures the penalization in terms of magnitude, hence obtaining BIC always increasing as the model becomes simpler; when this happens, even the BIC may favor very complex models. A possible mitigation might be provided by increasing the number of observations on the process, but it is known that sampling in high dimensions is rather difficult and computationally burdensome

In contrast, the adaptive reduction criteria here proposed appear to overcome this kind of issues. To provide some additional insights, let us consider the following problem: a 7-dimensional Gaussian mixture of size $n = 60$ is generated according to the procedure reported in section 10, for parameters $N_c = 10$, $\beta = 30$, $\delta_1 = 0.6$, $\delta_2 = 0.2$. From such a mixture, $N = 200000$ samples are drawn and a $n^a = 40$ components mixture is fitted by means of EM algorithm. This back and forth procedure is made in order to lose track of the original number of components, to add noise, and to provide a set of data which should be approximated sufficiently well with $n^b = 10$ components. A comparison between the adaptively chosen number and the BIC is reported in Fig 5.10.

Let us now consider the case of underfitting the data, that is a mixture of $n = 60$ $7-$*dimensional* components are generated out of $N_c = 15$ clusters for parameters $\beta = 0.6$, $\delta_1 = 0.6$, $\delta_2 = 0.2$; consequently, $N = 200000$ samples are drawn from such a mixture and $n^a = 10$ components are fitted by means of the EM algorithm over the samples. The results are reported in Fig. 5.11.

As it can be observed, both the BIC (expected, given the dimensionality of the problem) and the SMC suggest that no reduction should be made, hence a suitable model order is at least $n^b = 10$. As discussed, though, this kind of SMC curves appear when there is an underfitting of the data, and the model parameter number should be reconsidered to explore more complex representations.

A last experiment to investigate the scalability of such algorithm in high dimensional problems is reported. Let us consider $N = 200000$ samples

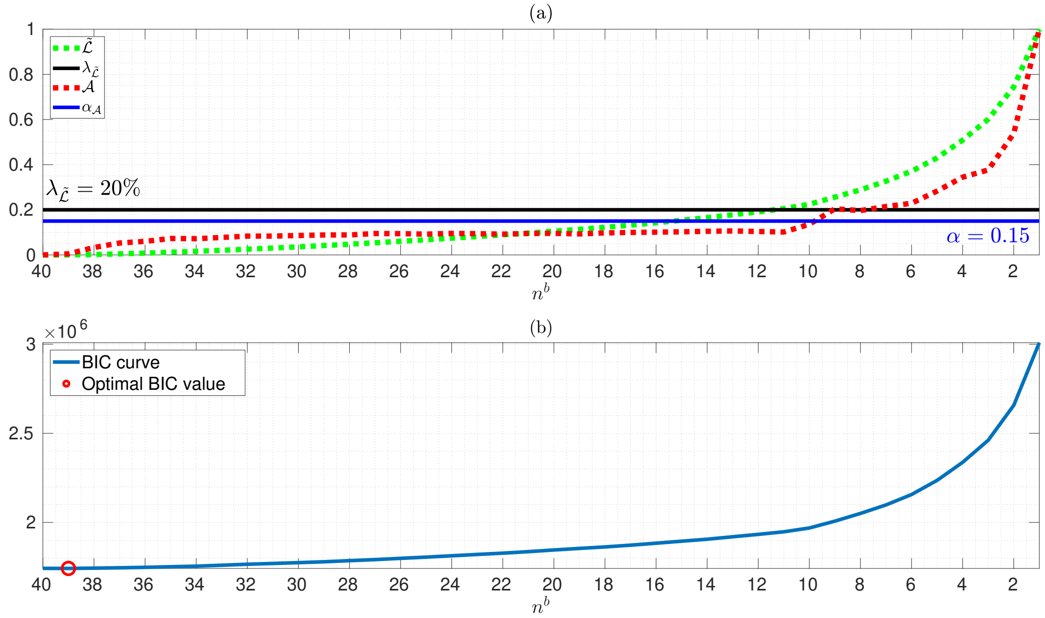

Figure 5.10: Set of $N = 200000$ samples fitted by means of EM algorithm with $n^a = 40$ Gaussian components. The adaptive reduction is employed to find a suitable model order. (a). Adaptive criteria curves (b). BIC curve.

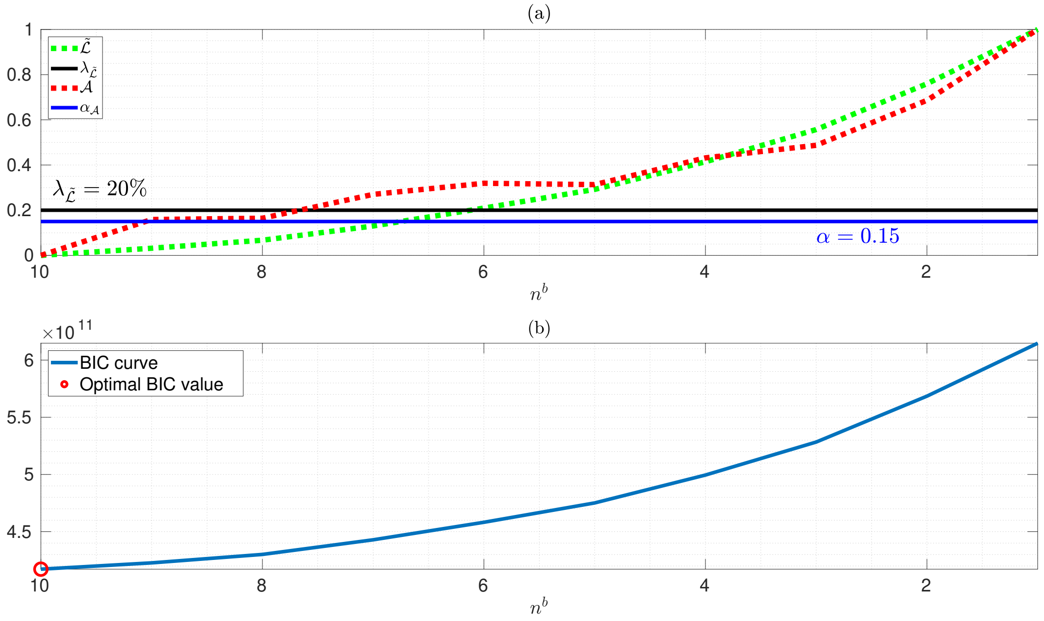

Figure 5.11: Set of $N = 200000$ samples fitted by means of EM algorithm with $n^a = 10$ Gaussian components. The adaptive reduction is employed to find a suitable model order. (a). Adaptive criteria curves (b). BIC curve.

generated from $n = 100$ 32-dimensional Gaussian components, regrouped in $N_c = 10$ clusters, with parameters $\beta = 60$ (to have a high probability of non-overlapping clusters), $\delta_1 = 0.4$ and $\delta_2 = 0.2$. The results are reported in Fig. 5.12.

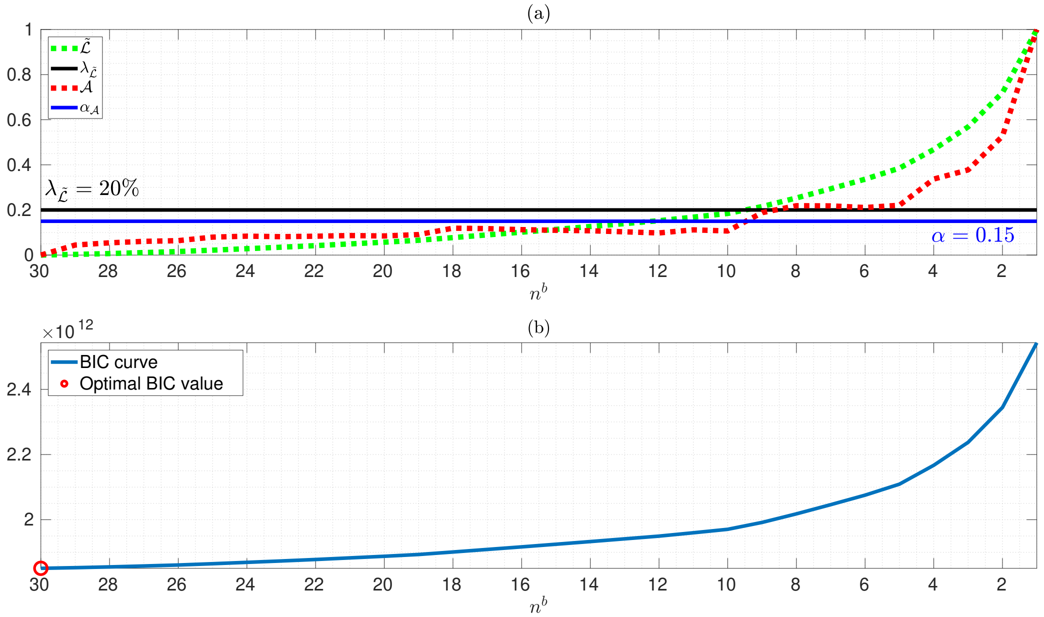

Figure 5.12: Set of $N = 200000$ samples fitted by means of EM algorithm with $n^a = 30$ Gaussian components. The adaptive reduction is employed to find a suitable model order. (a). Adaptive criteria curves (b). BIC curve.

Again, the SMC is able to correctly identify the number of clusters.

To conclude this section, there are few remarks which should be done. First of all, the coupling of EM and adaptive reduction seems to work really well in practice, even if not theoretically sound (soft-clustering vs hard-clustering). In any case, given its computational efficiency, it might represent a suitable algorithm for applications where resources are limited. The core idea, for which such an algorithm appears to perform well, relies on the fact that the adaptive reduction as the one presented in Sec. 4.4 aims to preserve the mixture modality; hence, even by detaching from the data, the final result still represents a good initialization for a possible subsequent instance of the EM algorithm. Moreover, as briefly mentioned, by looking at the corresponding curves $\tilde{\mathcal{L}}$ and $\mathcal{A}$, it is possible to figure out, visually, if the initial number of provided components may be sufficient to describe the data: significant increments for low compression ratios are in general symptom of underfitting models. Alternatively to the discussed algorithm, one could consider a modified version of the EM where, after each re-estimation of the components over the data, a reduction based on the ATC/SMC is performed. By recalling Fig. 4.3, if a merging does not have to take place, the adaptive algorithms will prevent such an event to happen. In other words, at each EM iteration, if there exist two (or more) components which do not introduce a significant increment in the $\tilde{\mathcal{L}}$ or $\mathcal{A}$ curves, then it is possible to perform a corresponding merging to reduce the model order.

Chapter 6

Conclusions and Future Work

In this dissertation, the topic of approximating mixture models to keep the complexity contained has been addressed in a bottom up approach. In Chapter 2, fundamentals have been reported in order to make the argumentation self-contained, and to provide a big picture about where the MRP arises and why it is important to address it rigorously. Many practical problems require a powerful, yet efficient, uncertainty representation in order to be approached robustly: mixture models can be a good candidate in that matter. Nonetheless, such representations are not issue-free, since in many applications the corresponding number of components can grow unbounded over time. In addition, even if the number of components is fixed a priori, it is non-trivial to select such a number in an *Occam's razor* perspective; using a considerably large number of components can lead to the phenomenon of overfitting, which, aside employing a rather significant amount of computational resources, may reduce the generalization capabilities of the model. In contrast, by selecting an insufficient number of components, the model can become particularly biased, and it may lose the ability to correctly predict/describe the underlying process.

In this regard, the topic of reducing a mixture complexity while finding a suitable number of components results to be an important, although non-trivial, task.

In Chapter 3, the MRP has been defined in an optimization perspective, where the dissimilarity between an original model and a corresponding simplified instance is wanted to be reduced while preserving representation accuracy. Such a problem, though, rarely admits a closed form solution, and it is often required to employ heuristics in order to address it. A common approach is to perform a greedy reduction of the components followed by a refinement phase. In the former, the mixture components are either merged together, by means of barycenter or BSDA, or pruned out, in an iterative manner. In the latter, the greedy reduction outcome serves as a starting point, and the corresponding parameters are refined by exploiting the original mixture information. Merging components together, though, can be done rarely in a closed form and often it is necessary to resort to numerical solutions as, for instance, fixed-point iteration algorithms. Even worse, the dissimilarity between mixtures can be computed in closed form for very few cases. For the above reasons, the MRP can be considered to be often analytically intractable. In the literature, such a fact has led to the employment of several dissimilarity measures in the same reduction pipeline, mostly for ease of computation, but this led to a lack of consistency in the solution (see 3.2.1, 5.1); from an optimization point of view, that is equivalent to minimize several different loss functions in the same problem, hence by yielding sub-optimal solutions.

Many of the mentioned issues can be solved by considering optimal transport theory. In Chapter 4, the problem is addressed by means of surrogate dissimilarities between mixtures, namely Composite Transportation Dissimilarities (CTDs), which are always available in closed form if the pairwise dissimilarities between mixture components can be computed. A CTD is the result of an optimal transport problem (OTP), which is a linear programming optimization problem. In such framework, the optimal merging action is the D-barycenter, which is rather easier to deal with if compared to the D-BSDA (see sec. 3.3); pruning is never optimal. Moreover, it is possible to obtain greedy reduction and refinement algorithms which are totally consistent with a single D-measure, and which result to be efficient in terms of computational resources; both the component reduction and subsequent refinement can be framed in a hard-clustering perspective, since all the actions taken in the OTT framework tend to aggregate similar components together and to recompute, eventually, the corresponding barycenters.

Furthermore, if the D_{FKL} is considered in such framework, a whole theory for adaptive reduction can be formulated, where the number of components is selected by the algorithm during the greedy descent. Such a criterion finds employment in several practical problems; for instance, in target tracking in clutter it can help to figure out how many objects may be present in the field of view, while filtering away the redundant components (see 5.2. In unsupervised learning problems, it can help to identify the number of classes/clusters in high dimensional spaces (see 5.3. A very good feature of the mentioned adaptive algorithm is that it results to be extremely efficient, hence suitable for real time applications which rely on very restricted computational resources.

To further investigate and prove the effectiveness of the proposed framework, in Chapter 5 are reported many numerical tests to investigate reliability of the reported algorithms and possible improvements/drawbacks; performing consistent reductions yields in general superior approximations in terms of the chosen D-measure, while maintaining a low computational burden if the OTT-based framework is considered (see sec. 5.1). In addition, the problem of reducing GGIW intensities is addressed in section 5.2; that kind of uncertainty representation is gaining a lot of attention in the context of extended target tracking, although it also suffers itself from the combinatorial explosion in the number of components if a Bayesian filtering setting is considered. GGIW hypotheses are called *product densities*, since are obtained as the product of simpler marginal distributions. When employed in target tracking problems, the corresponding reduction results to be more difficult if compared to the single marginal case, and having robust and accurate algorithms gains even more importance. Since the proposed reduction framework is general and can be applied both to mixtures and intensities, the GGIW intensity reduction can be directly addressed by exploiting the greedy reduction and refinement algorithms based on the OTT. As last analysis, in section 5.3 the adaptive reduction criteria are investigated in the model selection task for sample data; by at first overfitting the data, and by adaptively reducing the overly complex model, it is possible to provide a good trade-off between model complexity and accuracy, even for high dimensional problems, where other criteria like the AIC or BIC fail.
Overall, the main contributions of this dissertation are:

- to provide a general perspective on the whole mixture reduction pipeline in terms of optimal transport theory, mostly for the greedy reduction part, since the refinement scheme was already known in the literature; the whole work has focused mostly on the greedy reduction since, in the author's opinion, providing accurate initializations can help the refinement phase to achieve a fast convergence towards superior solutions, hence resulting to be a crucial point. In this regard, the proposed approach offers efficient, hybrid, consistent, greedy reduction algorithms for each given D-measure ([62], 4.3).

- To provide a debate on consistency, which is often neglected in many existing algorithms: numerical tests have proven why it is important to consider consistent pipelines when addressing the MRP. The formulation of the OTT-based reduction framework has been motivated mostly by such core topic, even because a reference algorithm/framework was missing in the literature to make fair[1] comparisons between new proposed solutions and existing approaches ([46], sec. 5.1).

- To embed an adaptive halting criterion in the greedy descent for the D_{FKL} case, hence to stop the reduction when a suitable number of components has been found. Such a method appears to work surprisingly well, in both the problems of target tracking and model selection for sample data. At the current state, it relies on the thresholding of some quantities, but the author is confident that such an operation can be removed in a near future ([80], sec. 4.4).

- To provide a comprehensive theory for GGIW intensity (adaptive) ([19], sec. 5.2), by showing how the problem can be addressed consistently with the D_{FKL} in an efficient, real-time suitable, manner.

- To investigate the employment of the adaptive reduction theory in the model selection task. From several experiments, it seems that the proposed criteria can outperform existing alternatives as the AIC and BIC, since they scale up well with dimensions (sec. 5.3).

- To provide an ensemble of FPI algorithms for the barycenter computation for many D-measures: this improved the efficiency of the overall OTT-based reduction framework, which hence does not have to rely on gradient descent optimization ([48], 3.3).

- To investigate the Likeness-based (LB) family of D-measures in both their analytical properties and peculiarities ([35], sec. 3.1.4, sec. 5.1).

Since the most encouraging result is the adaptive reduction criterion, as near future works the author would investigate further the effectiveness of such a method on real-world problems. Until now, all the data have been

[1] In the sense that many of the existing algorithms are either inconsistent or the corresponding performances are evaluated by a different D-measure than the one employed in the reduction. Given the proposed framework, the author hopes that future comparisons between new approaches will take into account all the argumentation here reported.

synthetically generated in order to have a ground truth for comparisons; since very good results have been obtained, the author expects to successfully employ the adaptive reduction on real world problems. As last minute finding, the author figured out that the SMC can be applied in practice to every D-measure listed in this work; nonetheless, at this stage, there are no theoretical guarantees as happens for the D_{FKL} case. Despite of this, several preliminary tests have shown that the SMC can be extended to any D-measure, even if lacking associativity of barycenters, and the corresponding results seem to stress out further how the features of a dissimilarity influence the MR outcome, both in the preserved mixture characteristics and the corresponding suitable number of components.

Although not reported in this work due to time and space constraints, the author has also addressed the topic of alternative surrogate functions to the CTDs (based on a regularization of the optimal transport problem), which also seem to perform interestingly when employed in the MRP. Nonetheless, the corresponding discussion would require a chapter on its own, hence it will probably be matter for future works.

Appendix A

Useful Formulas

A.1 List of D-measures

Here is reported a list of the main D-measures discussed in this work (all ratios between pdfs are assumed well defined):

$$\text{Kullback-Leibler} \quad D_{KL}(p\|q) = \int p(x)\log\frac{p(x)}{q(x)}dx, \tag{A.1}$$

$$\begin{aligned}\text{Jeffreys (symm-KL)} \quad D_{SKL}(p\|q) &= \frac{1}{2}(D_{KL}(p\|q) + D_{KL}(q\|p)) \\ &= \frac{1}{2}\int(p(x)-q(x))\log\frac{p(x)}{q(x)}\mathrm{d}x,\end{aligned} \tag{A.2}$$

$$\text{Skew Jeffreys}_{\alpha\in[0,1]} \quad D_J^{\alpha}(p\|q) = (1-\alpha)D_{KL}(p\|q) + \alpha D_{KL}(q\|p), \tag{A.3}$$

$$\underset{\text{aka Integral Square Error (ISE)}}{\text{Square L2 norm}} \quad D_{L2}(p\|q) = \int \big(p(x)-q(x)\big)^2\mathrm{d}x, \tag{A.4}$$

$$\text{Cauchy-Schwarz} \quad D_{CS}(p\|q) = -\log\left(\frac{\int p(x)q(x)\mathrm{d}x}{\sqrt{\big(\int p^2(x)\mathrm{d}x\big)\big(\int q^2(x)\mathrm{d}x\big)}}\right), \tag{A.5}$$

$$\text{Bhattacharyya distance} \quad D_B(p\|q) = -\log\int\sqrt{p(x)q(x)}\mathrm{d}x, \tag{A.6}$$

$$\text{Square Hellinger} \quad D_{H2}(p\|q) = \frac{1}{2}\int\Big(\sqrt{p(x)} - \sqrt{q(x)}\Big)^2 \mathrm{d}x \\ = 1 - \int \sqrt{p(x)q(x)}\mathrm{d}x, \tag{A.7}$$

$$\text{Pearson } \chi^2 \quad D_P(p\|q) = \frac{1}{2}\big(\int \frac{q(x)^2}{p(x)}\mathrm{d}x - 1\big), \tag{A.8}$$

$$\text{Neyman } \chi^2 \quad D_N(p\|q) = \frac{1}{2}\big(\int \frac{p(x)^2}{q(x)}\mathrm{d}x - 1\big), \tag{A.9}$$

$$\begin{array}{l}\text{Square}\\ \text{2-Wasserstein}\end{array} \quad D_{W2}(p\|q) = \inf_{\pi \in \mathcal{Q}_2} \iint \|x-y\|^2 \pi(x,y)dxdy \tag{A.10}$$

where $\mathcal{Q}_2$ is the set of all pdfs in $\mathbb{R}^d \times \mathbb{R}^d$ that have $p(x)$ and $q(y)$ as marginals and finite second order moments.

α-divergences

The Chernoff α-coefficient $c_\alpha(p,q)$, $\alpha \in (-\infty, \infty)$, defined as

$$c_\alpha(p,q) = \int p^\alpha(x) q^{1-\alpha}(x)\mathrm{d}x \tag{A.11}$$

allows to define two families of divergences: α-divergences of the I° and of then II° kind:

$$I^\circ \text{ kind } \alpha\text{-div} \qquad D'_\alpha(p\|q) = \frac{1}{\alpha(1-\alpha)}\big(1 - c_\alpha(p,q)\big), \tag{A.12}$$

$$II^\circ \text{ kind } \alpha\text{-div} \qquad D''_\alpha(p\|q) = -\log c_\alpha(p,q). \tag{A.13}$$

For some values of α, the α-divergences of the I° kind coincide with some of the previously listed divergences:

$$D'_\alpha(p\|q)\big|_{\alpha=-1} = D_P(p\|q) \qquad \text{Pearson } \chi^2 \tag{A.14}$$

$$\lim_{\alpha\to 0} D'_\alpha(p\|q) = D_{KL}(q\|p) \qquad \text{Reverse KL} \tag{A.15}$$

$$D'_\alpha(p\|q)\big|_{\alpha=0.5} = 4D_{H2}(p\|q) \qquad \text{Square Hellinger} \tag{A.16}$$

$$\lim_{\alpha\to 1} D'_\alpha(p\|q) = D_{KL}(p\|q) \qquad \text{Forward KL} \tag{A.17}$$

$$D'_\alpha(p\|q)\big|_{\alpha=2} = D_N(p\|q) \qquad \text{Neyman } \chi^2 \tag{A.18}$$

For $\alpha = 0.5$, $c_\alpha(p,q)$ coincides with the Bhattacharayya coefficient [88] and the corresponding α-divergence of the *II*° kind coincides with the Bhattacharayya distance.

A.2 Some identities

The *cross-likeness* $J^{i,j}$ between two Gaussians $\nu_i = \nu(x|\mu_i, \Sigma_i)$ and $\nu_j = \nu(x|\mu_j, \Sigma_j)$ is defined in [35] as

$$J^{i,j} = \int \nu_i(x) \cdot \nu_j(x) dx = \nu(\mu_i|\mu_j, \Sigma_i + \Sigma_j). \tag{A.19}$$

The *self-likeness* of a Gaussian ν_i is

$$J^{i,i} = \int \nu^2(x) dx = \nu(\mu_i|\mu_i, 2\Sigma_i) = |4\pi\Sigma_i|^{-\frac{1}{2}}. \tag{A.20}$$

If ν is without subscript, then the self-likeness is written $J^{\nu,\nu}$, so that $J^{\nu,\nu} = \nu(\mu|\mu, 2\Sigma)$, and the cross-likeness between ν_i and ν is written as $J^{i,\nu}$. The following hold true

$$\nu_i(x) \cdot \nu(x) = J^{i,\nu}\, \nu(x|\bar{\mu}_{i,\nu}, \bar{\Sigma}_{i,\nu}) \tag{A.21}$$

$$\text{where} \quad \begin{aligned} \bar{\mu}_{i,\nu} &= \left(\Sigma_i^{-1} + \Sigma^{-1}\right)^{-1}\left(\Sigma_i^{-1}\mu_i + \Sigma^{-1}\mu\right) \\ \bar{\Sigma}_{i,\nu} &= \left(\Sigma_i^{-1} + \Sigma^{-1}\right)^{-1} \end{aligned} \tag{A.22}$$

$$\nu^2(x) = J^{\nu,\nu}\, \nu\left(x\middle|\mu, \tfrac{1}{2}\Sigma\right) = \frac{1}{\sqrt{|4\pi\Sigma|}} \nu\left(x\middle|\mu, \tfrac{1}{2}\Sigma\right). \tag{A.23}$$

The partial derivatives of a Gaussian density $\nu = \nu(x|\mu, \Sigma)$ with respect to μ and Σ^{-1} are as follows:

$$\begin{aligned} \frac{\partial \nu}{\partial \mu} &= \Sigma^{-1}(x-\mu) \cdot \nu \\ \frac{\partial \nu}{\partial \Sigma^{-1}} &= \frac{1}{2}\left(\Sigma - (x-\mu)(x-\mu)^T\right) \cdot \nu. \end{aligned} \tag{A.24}$$

For any $\alpha \in (0,1)$ we have:

$$\begin{aligned} \nu^\alpha &= \frac{(2\pi)^{\frac{d}{2}(1-\alpha)}|\Sigma|^{\frac{1-\alpha}{2}}}{\alpha^{\frac{d}{2}}} \nu\left(x|\mu, \tfrac{1}{\alpha}\Sigma\right), \\ \nu^{1-\alpha} &= \frac{(2\pi)^{\frac{d}{2}\alpha}|\Sigma|^{\frac{\alpha}{2}}}{(1-\alpha)^{\frac{d}{2}}} \nu\left(x|\mu, \tfrac{1}{1-\alpha}\Sigma\right). \end{aligned} \tag{A.25}$$

The Chernoff α-coefficient (A.11) between ν_i and ν_j is

$$c_\alpha(\nu_i,\nu_j) = \left(\frac{|\bar{\Sigma}^\alpha_{i,j}|}{|\Sigma_i|^\alpha |\Sigma_j|^{1-\alpha}}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(\mu_i-\mu_j)^T(\widetilde{\Sigma}^\alpha_{i,j})^{-1}(\mu_i-\mu_j)} \tag{A.26}$$

$$\text{so that} \quad \nu_i^\alpha \cdot \nu_j^{1-\alpha} = c_\alpha(\nu_i,\nu_j)\cdot \nu(x|\bar{\mu}^\alpha_{i,j}, \bar{\Sigma}^\alpha_{i,j}), \tag{A.27}$$

$$\begin{aligned} & \bar{\Sigma}^\alpha_{i,j} = \left(\alpha\Sigma_i^{-1} + (1-\alpha)\Sigma_j^{-1}\right)^{-1}, \\ \text{where} \quad & \widetilde{\Sigma}^\alpha_{i,j} = \tfrac{1}{\alpha}\Sigma_i + \tfrac{1}{1-\alpha}\Sigma_j, \\ & \bar{\mu}^\alpha_{i,j} = \bar{\Sigma}^\alpha_{i,j}\left(\alpha\Sigma_i^{-1}\mu_i + (1-\alpha)\Sigma_j^{-1}\mu_j\right). \end{aligned} \tag{A.28}$$

By exploiting (A.24), the derivatives of ν^α and $\nu^{1-\alpha}$ w.r.t. the parameters (μ, Σ^{-1}) can be obtained in a straightforward manner.
Similarly, using (A.24) the derivatives of the cross-likeness $J^{i,\nu}$, between ν_i and ν and of the self-likeness $J^{\nu,\nu}$, with respect to the parameters (μ, Σ^{-1}) of ν can be obtained:

$$\begin{aligned} &\frac{\partial J^{i,\nu}}{\partial \mu} = \Sigma^{-1}(\bar{\mu}_{i,\nu} - \mu)J^{i,\nu}, \qquad \frac{\partial J^{\nu,\nu}}{\partial \mu} = 0 \\ &\frac{\partial J^{i,\nu}}{\partial \Sigma^{-1}} = \frac{1}{2}\big(\Sigma - (\bar{\Sigma}_{i,\nu} + (\bar{\mu}_{i,\nu} - \mu)(\bar{\mu}_{i,\nu} - \mu)^T)\big)J^{i,\nu} \\ &\frac{\partial J^{\nu,\nu}}{\partial \Sigma^{-1}} = \frac{1}{2}J^{\nu,\nu}\Sigma \end{aligned} \tag{A.29}$$

Analogous derivatives can be obtained for the Chernoff α-coefficient, omitted due to their cumbersome forms.

Appendix B

Proofs

All the following material represents, for the most part, unpublished contributions.

B.1 Equivalence of BSDAs and barycenters

B.1.1 D_{FKL}-BSDA and D_{FKL}-barycenter equivalence

Proof. Let us consider a mixture of densities $p = \boldsymbol{w}^T \boldsymbol{q} = \sum_{i=1}^n w_i q_i$, $p \in \mathcal{Q}_{\text{mix}}$, and a generic distribution q. It follows that:

$$\begin{aligned}
q^* &= \arg\min_{q\in\mathcal{Q}} D_{FKL}(\boldsymbol{w}^T\boldsymbol{q}\|q) = \arg\min_{q\in\mathcal{Q}} \int \sum_{i=1}^n w_i q_i \log \frac{\sum_{i=1}^n w_i q_i}{q} \mathrm{d}x = \\
&= \arg\min_{q\in\mathcal{Q}} \Bigg[\underbrace{\int \sum_{i=1}^n w_i q_i \log \sum_{i=1}^n w_i q_i \mathrm{d}x}_{\perp\!\!\!\perp q} - \int \sum_{i=1}^n w_i q_i \log q \mathrm{d}x \Bigg] = \\
&= \arg\min_{q\in\mathcal{Q}} \Bigg[\underbrace{\sum_{i=1}^n w_i \int q_i \log q_i \mathrm{d}x}_{\perp\!\!\!\perp q} - \sum_{i=1}^n w_i \int q_i \log q \mathrm{d}x \Bigg] = \qquad \text{(B.1)} \\
&= \arg\min_{q\in\mathcal{Q}} \sum_{i=1}^n w_i \Bigg[\int q_i \log q_i \mathrm{d}x - \int q_i \log q \mathrm{d}x \Bigg] = \\
&= \arg\min_{q\in\mathcal{Q}} \sum_{i=1}^n w_i D_{FKL}(q_i\|q) = \arg\min_{q\in\mathcal{Q}} m_{D_{FKL}}(q|\boldsymbol{w},\boldsymbol{q}) = \hat{q}
\end{aligned}$$

□

B.1.2 D_{L2}-BSDA and D_{L2}-barycenter equivalence

Proof. Let us consider a mixture of densities $p = \boldsymbol{w}^T\boldsymbol{q} = \sum_{i=1}^n w_i q_i$, $p \in \mathcal{Q}_{\text{mix}}$, and a generic distribution q. It follows that:

$$\begin{aligned}
q^* &= \arg\min_{q\in\mathcal{Q}} D_{L2}(\boldsymbol{w}^T\boldsymbol{q}\|q) = \arg\min_{q\in\mathcal{Q}} \int \big(\boldsymbol{w}^T\boldsymbol{q} - q\big)^2 \mathrm{d}x = \\
&= \arg\min_{q\in\mathcal{Q}} \Bigg[\underbrace{\int \big(\sum_{i=1}^n w_i q_i\big)^2 \mathrm{d}x}_{\perp\!\!\!\perp q} - 2\sum_{i=1}^n w_i \int q_i q \mathrm{d}x + \int q^2 \mathrm{d}x \Bigg] = \\
&= \arg\min_{q\in\mathcal{Q}} \Bigg[\underbrace{\sum_{i=1}^n w_i \int q_i^2 \mathrm{d}x}_{\perp\!\!\!\perp q} - 2\sum_{i=1}^n w_i \int q_i q \mathrm{d}x + \underbrace{\sum_{i=1}^n w_i}_{=1} \int q^2 \mathrm{d}x \Bigg] = \qquad \text{(B.2)} \\
&= \arg\min_{q\in\mathcal{Q}} \sum_{i=1}^n w_i \Bigg[\int q_i^2 \mathrm{d}x - 2\int q_i q \mathrm{d}x + \int q^2 \mathrm{d}x \Bigg] = \\
&= \arg\min_{q\in\mathcal{Q}} \sum_{i=1}^n w_i D_{L2}(q_i\|q) = \arg\min_{q\in\mathcal{Q}} m_{D_{L2}}(q|\boldsymbol{w},\boldsymbol{q}) = \hat{q}
\end{aligned}$$

□

B.2 D_{FKL}-barycenter properties

B.2.1 Uniqueness of D_{FKL}-barycenters for the exponential family

D_{FKL}-barycenter uniqueness for the exponential family. From the definition of D_{FKL}, exploiting the expression (2.45) of pdfs in the exponential family, the following formula is easily derived:

$$\begin{aligned} m_{D_{FKL}}(q(\eta)|\boldsymbol{w},\boldsymbol{q}) &= \sum_{i=1}^{n} w_i \big(\eta_i^T b(\eta_i) - a(\eta_i)\big) - \sum_{i=1}^{n} w_i \big(\eta^T b(\eta_i) - a(\eta)\big) \\ &= \sum_{i=1}^{n} w_i \big(\eta_i^T b(\eta_i) - a(\eta_i)) - \eta^T \Big(\sum_{i=1}^{n} w_i b(\eta_i)\Big) + \bar{w}\, a(\eta). \end{aligned} \tag{B.3}$$

Defining the function $f(\eta) = m_{D_{FKL}}(q(\eta)|\boldsymbol{w},\boldsymbol{q})$, it is easy to see that its Hessian is proportional to the Hessian of $a(\eta)$ via the positive coefficient $\bar{w}$, i.e. $d^2f/d\eta^2 = \bar{w}(d^2a/d\eta^2)$, and therefore is positive definite, from (2.49). It follows that $f(\eta)$ has a unique minimum $\hat{\eta}$ at which the gradient $df/d\eta$ vanishes

$$\exists!\, \hat{\eta} \in \Lambda : \ \frac{df(\eta)}{d\eta}\Big|_{\hat{\eta}} = 0 \quad \Leftrightarrow \quad -\sum_{i=1}^{n} w_i b(\eta_i) + \bar{w} \left(\frac{da(\eta)}{d\eta}\right)^T_{\hat{\eta}} = 0. \tag{B.4}$$

Recalling that $(da/d\eta)^T = b(\eta)$ (the first of the two properties (2.49)), the thesis, eq. (3.88), follows. □

B.2.2 D_{FKL}-barycenter associativity for the exponential family

Associativity of D_{FKL}-barycenters for the exponential family. This result is a direct consequence of equation (3.88), that characterizes the parameter $\hat{\eta}$ of the barycenter of a weighted set $(\boldsymbol{w},\boldsymbol{q}) = \{w_i q_i\}_{i=1}^n$ of distributions $q_i \in \mathcal{Q}$. The proof is obtained considering a pair of disjoint subsets $\mathcal{I}_1$ and $\mathcal{I}_2$ of the interval $[1\!:\!n]$. Associativity is proved by showing that the identity (3.73), which defines the associativity property, is true. Let $\hat{\eta}_1$ and $\hat{\eta}_2$ denote the natural parameters of the barycenters of the weighted subsets of $(\boldsymbol{w},\boldsymbol{q})$ associated to the subsets of indices $\mathcal{I}_1$ and $\mathcal{I}_2$, and let $\hat{\eta}_{1,2}$ denote the natural parameter of the barycenter of the weighted subset of $(\boldsymbol{w},\boldsymbol{q})$ associated to the subsets of indices $\mathcal{I}_1 \cup \mathcal{I}_2$. To prove associativity, one must show that

$$\begin{aligned} &q(\hat{\eta}_{1,2}) = \bar{\Phi}_{D_{FKL}}\big(\{\bar{w}_{\mathcal{I}_1} q(\hat{\eta}_1), \bar{w}_{\mathcal{I}_2} q(\hat{\eta}_2)\}\big), \\ &\text{where} \quad \bar{w}_{\mathcal{I}_j} = \sum_{i\in\mathcal{I}_j} w_i, \quad j \in \{1,2\}. \end{aligned} \tag{B.5}$$

According to equation (3.88), $\hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_{1,2}$ are such that

$$\begin{aligned}
\hat{\eta}_1 : \quad & \bar{w}_{\mathcal{I}_1} b(\hat{\eta}_1) = \sum_{i \in \mathcal{I}_1} w_i \, b(\eta_i), \\
\hat{\eta}_2 : \quad & \bar{w}_{\mathcal{I}_2} b(\hat{\eta}_2) = \sum_{i \in \mathcal{I}_2} w_i \, b(\eta_i), \\
\hat{\eta}_{1,2} : \quad & \bar{w}_{\mathcal{I}_1 \cup \mathcal{I}_2} b(\hat{\eta}_{1,2}) = \sum_{i \in \mathcal{I}_1 \cup \mathcal{I}_2} w_i \, b(\eta_i).
\end{aligned} \tag{B.6}$$

The proof is easily obtained by considering that

$$\begin{aligned}
\bar{w}_{\mathcal{I}_1 \cup \mathcal{I}_2} =& \bar{w}_{\mathcal{I}_1} + \bar{w}_{\mathcal{I}_2}, \\
\sum_{i \in \mathcal{I}_1 \cup \mathcal{I}_2} w_i \, b(\eta_i) =& \sum_{i \in \mathcal{I}_1} w_i \, b(\eta_i) + \sum_{i \in \mathcal{I}_2} w_i \, b(\eta_i),
\end{aligned} \tag{B.7}$$

so that the condition (B.6) on $\hat{\eta}_{1,2}$ can be rewritten as

$$\hat{\eta}_{1,2} : \ \big(\bar{w}_{\mathcal{I}_1} + \bar{w}_{\mathcal{I}_2}\big) b(\hat{\eta}_{1,2}) = \bar{w}_{\mathcal{I}_1} \, b(\hat{\eta}_1) + \bar{w}_{\mathcal{I}_2} \, b(\hat{\eta}_2), \tag{B.8}$$

which is the condition on the barycenter of the weighted set $\{\bar{w}_{\mathcal{I}_j} q(\hat{\eta}_j)\}_{j \in \{1,2\}}$, and this concludes the proof. □

B.2.3 D-measure joint convexity and CTD upper bounds

Joint convexity and upper bounds. Let us consider four distributions $p_1, p_2, q_1, q_2 \in \mathbb{R}^d$ and a coefficient $\alpha \in [0,1]$; a D-measure is said to be jointly convex (j.c.) in the arguments if it holds that:

$$D(\alpha p_1 + (1-\alpha) p_2 \| \alpha q_1 + (1-\alpha) q_2) \leq \alpha D(p_1 \| q_1) + (1-\alpha) D(p_2 \| q_2) \tag{B.9}$$

Given a jointly convex D-measure and two mixture densities p^a and p^b, let W be a transportation plan between p^a and p^b for which one can write $p^a = \sum_{i=1}^{n^a} w_i^a q_i^a = \sum_{i=1}^{n^a} \sum_{j=1}^{n^b} W_{i,j} q_i^a$ and $p^b = \sum_{j=1}^{n^b} w_i^b q_i^b = \sum_{j=1}^{n^b} \sum_{i=1}^{n^a} W_{i,j} q_j^b$;

then it holds the following result:

$$\begin{aligned}D(p^a\|p^b) = D\big(\sum_{i=1}^{n^a} w_i^a q_i^a \| \sum_{j=1}^{n^b} w_j^b q_j^b\big) = \\ = D\big(\sum_{i=1}^{n^a}\sum_{j=1}^{n^b} W_{i,j} q_i^a \| \sum_{j=1}^{n^b}\sum_{i=1}^{n^a} W_{i,j} q_j^b\big) \leq \\ \underbrace{\leq}_{\text{j.c.}} \sum_{i=1}^{n^a}\sum_{j=1}^{n^b} W_{i,j} D(q_i^a\|q_j^b)\end{aligned} \tag{B.10}$$

Of course, if $W = \widehat{W}$, that is the optimal transportation plan is considered, it holds that:

$$D(p^a\|p^b) \leq C_D(p^a\|p^b) \tag{B.11}$$

□

B.2.4 B_D, C_D and C_D^R properties

In this section, additional properties, propositions and theorems are reported to prove the results presented in Sec. 4.3.

Proposition B.2.1. *Consider nondegenerate mixtures p^a, of size n^a, and p^b, of size $n^b < n^a$, such that do not exists any component q_i^a of p^a with equal divergence from any two components of p^b, i.e.*

$$\forall i \in [1\!:\!n^a], \quad D(q_i^a\|q_h^b) \neq D(q_i^a\|q_k^b), \quad \forall h,k \in [1\!:\!n^b]. \tag{B.12}$$

Then, the optimal relaxed transportation plan $\breve{W} \in \breve{\mathcal{T}}(\boldsymbol{w}^a)$, solution of (4.10), *is a sparse matrix, such that*

$$\forall i \in [1\!:\!n^a], \quad \begin{aligned}\breve{W}_{i,j^*} &= w_i^a, \quad j^* = \operatorname*{arg\,min}_{j\in[1:n^b]} \big(D(q_i^a\|q_j^b)\big) \\ \breve{W}_{i,j} &= 0, \qquad j \neq j^*.\end{aligned} \tag{B.13}$$

Proof. Rewriting (4.13)

$$C_D^R(p^a\|p^b) = \min_{W\in\breve{\mathcal{T}}(\boldsymbol{w}^a)} \Big(\sum_{i=1}^{n^a}\sum_{j=1}^{n^b} W_{i,j} D(q_i^a\|q_j^b)\Big), \tag{B.14}$$

the optimal relaxed transportation plan $\breve{W}$ solves

$$\min_{W \in \breve{\mathcal{T}}(\boldsymbol{w}^a)} \Big(\sum_{i=1}^{n^a} \sum_{j=1}^{n^b} W_{i,j} D(q_i^a \| q_j^b) \Big) = \sum_{i=1}^{n^a} \left(\min_{W_{i,:} \in \breve{\mathcal{T}}_i(\boldsymbol{w}^a)} \sum_{j=1}^{n^b} W_{i,j} D(q_i^a \| q_j^b) \right), \tag{B.15}$$

where $W_{i,:}$ denotes the i-th row of the matrix W, and $\breve{\mathcal{T}}_i(\boldsymbol{w}^a) = \{W_{i,:} \in \mathbb{R}_+^{1 \times n^b} : W_{i,:}\mathbb{1}_{n^b} = w_i^a\}$. It follows that the optimal relaxed transportation plan $\breve{W}$ can be independently computed row by row

$$i \in [1 \colon n^a], \quad \breve{W}_{i,:} = \arg \left(\min_{W_{i,:} \in \breve{\mathcal{T}}_i(\boldsymbol{w}^a)} \sum_{j=1}^{n^b} W_{i,j} D(q_i^a \| q_j^b) \right). \tag{B.16}$$

The transportation plan $\breve{W}$ defined in (B.13) is such that

$$\sum_{j=1}^{n^b} \breve{W}_{i,j} D(q_i^a \| q_j^b) = w_i^a D(q_i^a \| q_{j^*}^b) \leq \sum_{j=1}^{n^b} W_{i,j} D(q_i^a \| q_j^b), \quad \forall W_{i,:} \in \breve{\mathcal{T}}_i(\boldsymbol{w}^a), \tag{B.17}$$

because by assumption (B.12)

$$D(q_i^a \| q_{j^*}^b) < D(q_i^a \| q_j^b), \quad \forall j \neq j^*. \tag{B.18}$$

Equality in (B.17) happens only for $W_{i,:} = \breve{W}_{i,:}$, This concludes the proof. □

Lemma B.2.2. *Consider a D-measure such that the D-barycenter of any given mixture exists and is unique. Given a nondegenerate mixture $p = \boldsymbol{w}^T \boldsymbol{q} \in \mathcal{Q}_{mix}$ with n components $\{w_i q_i\}_{i=1}^n$, with all weights w_i strictly positive, the following inequalities hold true*

$$\forall i, j \in [1 \colon n] > 0, \quad \begin{array}{l} w_i D(q_i \| \hat{q}_{i,j}) + w_j D(q_j \| \hat{q}_{i,j}) < w_i D(q_i \| q_j) \\ w_i D(q_i \| \hat{q}_{i,j}) + w_j D(q_j \| \hat{q}_{i,j}) < w_j D(q_j \| q_i), \end{array} \tag{B.19}$$

where $\hat{q}_{i,j} = \bar{\Phi}_D(w_i q_i + w_j q_j)$. Moreover

$$D(q_i \| \hat{q}_{i,j}) < D(q_i \| q_j), \quad D(q_j \| \hat{q}_{i,j}) < D(q_j \| q_i). \tag{B.20}$$

Proof. The proof is obtained by exploiting the uniqueness of the barycenter

$$w_i D(q_i \| \hat{q}_{i,j}) + w_j D(q_j \| \hat{q}_{i,j}) < w_i D(q_i \| q) + w_j D(q_j \| q), \quad \forall q \in \mathcal{Q} \setminus \{\hat{q}_{i,j}\} \tag{B.21}$$

Replacing in (B.21) $q = q_i$ we get the first of (B.19), while the second is obtained replacing $q = q_j$. The inequalities (B.20) are a simple consequence of (B.19). □

Remark 3. *If the D-barycenters are not unique, then inequalities* (B.19) *and* (B.20) *of Lemma B.2.2 hold true if the strict inequality $<$ is replaced by the non-strict inequality $\leq$.*

Proposition B.2.3. *Given a mixture $p = \boldsymbol{w}^T \boldsymbol{q} \in \mathcal{Q}_{mix}$ with n components $\{w_i q_i\}_{i=1}^n$, with all weights w_i strictly positive, let $i, j, h \in [1:n]$ be three indices such that*

$$\begin{aligned} w_i D(q_i \| \hat{q}_{i,j}) + w_j D(q_j \| \hat{q}_{i,j}) &< w_i D(q_i \| \hat{q}_{i,h}) + w_h D(q_h \| \hat{q}_{i,h}) \\ w_i D(q_i \| \hat{q}_{i,j}) + w_j D(q_j \| \hat{q}_{i,j}) &< w_j D(q_j \| \hat{q}_{j,h}) + w_h D(q_h \| \hat{q}_{j,h}), \end{aligned} \tag{B.22}$$

where $\hat{q}_{i,j} = \bar{\Phi}_D(w_i q_i + w_j q_j)$. Then

$$\begin{aligned} D(q_i \| \hat{q}_{i,j}) &< D(q_i \| q_h) \\ D(q_j \| \hat{q}_{i,j}) &< D(q_j \| q_h). \end{aligned} \tag{B.23}$$

Proof. From Lemma B.2.2 and assumption (B.22) we get

$$\begin{aligned} w_i D(q_i \| \hat{q}_{i,j}) + w_j D(q_j \| \hat{q}_{i,j}) &< w_i D(q_i \| q_h) \\ w_i D(q_i \| \hat{q}_{i,j}) + w_j D(q_j \| \hat{q}_{i,j}) &< w_j D(q_j \| q_h). \end{aligned} \tag{B.24}$$

Since $w_j D(q_j \| \hat{q}_{ij}) > 0$, from the first of (B.24) the first of (B.23) follows, and since $w_i D(q_i \| \hat{q}_{ij}) > 0$, from the second of (B.24) the second of (B.23) follows. □

Theorem B.2.4 (Equality between B_D, C_D and C_D^R)**.** *Given a mixture $p = \boldsymbol{w}^T \boldsymbol{q} \in \mathcal{Q}_{mix}$ with n components, $\{w_i q_i\}_{i=1}^n$, $n > 2$, with all weights w_i strictly positive, let $\{i^*, j^*\} \subset [1:n]$, $i^* \neq j^*$, be a pair of indices such that $\forall r \in [1:n] \setminus \{i^*, j^*\}$, the following inequalities are true*

$$\begin{aligned} w_{i^*} D(q_{i^*} \| \hat{q}_{i^*,j^*}) + w_{j^*} D(q_{j^*} \| \hat{q}_{i^*,j^*}) &\leq w_{i^*} D(q_{i^*} \| \hat{q}_{i^*,r}) + w_r D(q_r \| \hat{q}_{i^*,r}) \\ w_{i^*} D(q_{i^*} \| \hat{q}_{i^*,j^*}) + w_{j^*} D(q_{j^*} \| \hat{q}_{i^*,j^*}) &\leq w_{j^*} D(q_{j^*} \| \hat{q}_{j^*,r}) + w_r D(q_r \| \hat{q}_{j^*,r}), \end{aligned} \tag{B.25}$$

where, as usual, $\hat{q}_{i^*,j^*} = \bar{\Phi}_D(w_{i^*}q_{i^*} + w_{j^*}q_{j^*})$*. Let*

$$\tilde{p}_{i^*,j^*} = p - (w_{i^*}q_{i^*} + w_{j^*}q_{j^*}) + (w_{i^*} + w_{j^*})\hat{q}_{i^*,j^*}. \tag{B.26}$$

Then, $C_D(p\|\tilde{p}_{i^*,j^*})$ *(see eq.* (4.4)*) and* $C_D^R(p\|\tilde{p}_{i^*,j^*})$ *(see eq.* (4.13)*) coincide and are as follows:*

$$C_D(p\|\tilde{p}_{i^*,j^*}) = C_D^R(p\|\tilde{p}_{i^*,j^*}) = w_{i^*}D(q_{i^*}\|\hat{q}_{i^*,j^*}) + w_{j^*}D(q_{j^*}\|\hat{q}_{i^*,j^*}). \tag{B.27}$$

Proof. To make simpler the notation, without loss of generality, in the proof we assume that $i^* = n-1$, and $j^* = n$, so that assumptions (B.25) are rewritten as:

$$\begin{aligned}
&w_{n-1}D(q_{n-1}\|\hat{q}_{n-1,n}) + w_nD(q_n\|\hat{q}_{n-1,n}) \le w_{n-1}D(q_{n-1}\|\hat{q}_{n-1,r}) + w_rD(q_r\|\hat{q}_{n-1,r})\\
&w_{n-1}D(q_{n-1}\|\hat{q}_{n-1,n}) + w_nD(q_n\|\hat{q}_{n-1,n}) \le w_nD(q_n\|\hat{q}_{n,r}) + w_rD(q_r\|\hat{q}_{n,r}),
\end{aligned} \tag{B.28}$$

for all $r \in [0\!:\!n-2]$, (B.26) becomes

$$\tilde{p}_{n-1,n} = p - (w_{n-1}q_{n-1} + w_nq_n) + (w_{n-1} + w_n)\hat{q}_{n-1,n}, \tag{B.29}$$

and the thesis (B.27) becomes

$$C_D(p\|\tilde{p}_{n-1,n}) = C_D^R(p\|\tilde{p}_{n-1,n}) = w_{n-1}D(q_{n-1}\|\hat{q}_{n-1,n}) + w_nD(q_n\|\hat{q}_{n-1,n}). \tag{B.30}$$

Let $\boldsymbol{D} \in \mathbb{R}_+^{n\times(n-1)}$ be the cost matrix, whose components are as follows

$$\begin{aligned}
\boldsymbol{D}_{i,j} &= D(q_i\|q_j), &&\forall i \in [1:n],\ j \in [1:n-2],\\
\boldsymbol{D}_{i,n-1} &= D(q_i\|\hat{q}_{n-1,n}), &&\forall i \in [1:n].
\end{aligned} \tag{B.31}$$

The structure of a relaxed optimal transportation plan $\breve{W}$ associated to the $C_D^R(p\|\tilde{p}_{n-1,n})$ is investigated below (the general structure has been already investigated in Proposition B.2.1).

The cost minimized by $\breve{W}$ is

$$\langle W, \boldsymbol{D}\rangle = \sum_{i=1}^{n}\Big(W_{i,n-1}D(q_i\|\hat{q}_{n-1,n}) + \sum_{j=1}^{n-2}W_{i,j}D(q_i\|q_j)\Big). \tag{B.32}$$

In the RCTD computation problem there are n independent constraints on the rows of the relaxed feasible transportation matrices $W \in \mathbb{R}_+^{n\times(n-1)}$: a feasible $W \in \breve{\mathcal{T}}(\boldsymbol{w})$ must be such that $W_{i,:}\mathbb{1}_{n-1} = w_i$, $i = 1, \dots, n$. Then, the problem can be written as

$$\begin{aligned} C_D^R(p|\tilde{p}_{n-1,n}) &= \min_{W \in \breve{\mathcal{T}}(\boldsymbol{w})} \sum_{i=1}^{n} \left(W_{i,n-1} D(q_i \| \hat{q}_{n-1,n}) + \sum_{j=1}^{n-2} W_{i,j} D(q_i \| q_j) \right) \\ &= \sum_{i=1}^{n} \min_{W_{i,:}\mathbb{1}_{n-1} = w_i} \Big(W_{i,n-1} D(q_i \| \hat{q}_{n-1,n}) + \sum_{j=1}^{n-2} W_{i,j} D(q_i \| q_j) \Big). \end{aligned} \tag{B.33}$$

Note that $D(q_i \| q_i) = 0$, for $i = 1, \dots, n-2$. Thus, the first $n-2$ rows of the optimal relaxed transportation plan $\breve{W}$ are:

$$\breve{W}_{i,j} = 0, \text{ for } i \neq j, \quad \breve{W}_{i,i} = w_i, \quad i = 1, \dots, n-2, \tag{B.34}$$

so that in (B.33) all terms with $i = 1, \dots, n-2$ vanish, and we get

$$\begin{aligned} C_D^R(p|\tilde{p}_{n-1,n}) = &\min_{W_{n-1,:}\mathbb{1}_{n-1} = w_{n-1}} \Big(W_{n-1,n-1} D(q_{n-1} \| \hat{q}_{n-1,n}) + \sum_{j=1}^{n-2} W_{n-1,j} D(q_{n-1} \| q_j) \Big) \\ &+ \min_{W_{n,:}\mathbb{1}_{n-1} = w_n} \Big(W_{n,n-1} D(q_n \| \hat{q}_{n-1,n}) + \sum_{j=1}^{n-2} W_{n,j} D(q_n \| q_j) \Big). \end{aligned} \tag{B.35}$$

Thanks to assumptions (B.25), rewritten as: (B.28), and to Proposition B.2.3 (actually, its version with non strict inequalities), we have

$$\begin{aligned} D(q_{n-1} \| \hat{q}_{n-1,n}) \le D(q_{n-1} \| q_j) \\ D(q_n \| \hat{q}_{n-1,n}) \le D(q_n \| q_j) \end{aligned} \quad \forall j \in [1\!:\!n]. \tag{B.36}$$

It follows that the last two rows of an optimal transportation plan $\breve{W} \in \breve{\mathcal{T}}(\boldsymbol{w})$ can be chosen as follows:

$$\begin{aligned} \breve{W}_{n-1,j} = 0, \text{ for } j \in [1\!:\!n-2], \quad \breve{W}_{n-1,n-1} = w_{n-1}, \\ \breve{W}_{n,j} = 0, \text{ for } j \in [1\!:\!n-2], \quad \breve{W}_{n,n-1} = w_n, \end{aligned} \tag{B.37}$$

(this choice for an optimal transportation plan $\breve{W}$ is univocal if strict inequalities are assumed in (B.25), so that inequalities (B.36) are strict, too), so that

$$C_D^R(p \| \tilde{p}_{n-1,n}) = w_{n-1} D(q_{n-1} \| \hat{q}_{n-1,n}) + w_n D(q_n \| \hat{q}_{n-1,n}), \tag{B.38}$$

which is identity of the two rightmost terms of the thesis (B.30). To prove the identity of the first two terms ($C_D(p\|\tilde{p}_{n-1,n}) = C_D^R(p\|\tilde{p}_{n-1,n})$ it is sufficient to see that the computed optimal relaxed transportation plan is feasible also for the non-relaxed transportation problem, i.e. $\widehat{W} \in \mathcal{T}(\boldsymbol{w}, \tilde{\boldsymbol{w}}) \subset \breve{\mathcal{T}}(\boldsymbol{w})$, where $\tilde{\boldsymbol{w}} \in \Delta^{n-2} \subset \mathbb{R}_+^{n-1}$ is the set of weights associated to $\tilde{p}_{n-1,n}$, i.e. , $\tilde{w}_i = w_i$, $i \in [1\!:\!n-2]$ and $\tilde{w}_{n-1} = w_{n-1} + w_n$. Indeed, $\mathbb{1}_n^T \widehat{W} = \tilde{w}$. □

At this point, it is left to prove that the merging choice associated to the optimal bound $\breve{B}_D$ is the one which introduces also the minimum CTD between the mixture before and the one after merging. To ease the corresponding discussion, few more definitions and remarks have to be reported.

Definition B.2.1. (Child and parent mixtures) *At first, recall that the superscript $(\cdot)$ represents the order of a given mixture or, alternatively, the number of elements contained in a set; moreover, the quantities corresponding to such mixture/set will possess such a superscript to be labelled as belonging the considered entity. Thus said, consider a mixture $p^{(n)} = (\boldsymbol{w}^{(n)})^T \boldsymbol{q}^{(n)}$ of size n, with components $q_i^{(n)} \in \mathcal{Q}$, and consider a partition $\mathcal{I}^{(m)}$ of size m of the interval $[1\!:\!n]$ (of course, $m < n$). The* child mixture of $p^{(n)}$ induced by the partition $\mathcal{I}^{(m)}$ *is the mixture $p^{(m)} = (\boldsymbol{w}^{(m)})^T \boldsymbol{q}^{(m)}$ of size m defined as follows:*

$$\begin{aligned} &\boldsymbol{w}^{(m)} \in \Delta^{m-1}: \quad w_j^{(m)} = \bar{w}^{(n)}_{\mathcal{I}_j^{(m)}} = \sum_{i \in \mathcal{I}_j^{(m)}} w_i^{(n)}, \\ & && j \in [1\!:\!m]. \qquad \text{(B.39)} \\ &\boldsymbol{q}^{(m)} \in (\mathcal{Q})^{(m)}: \quad q_j^{(m)} = \hat{q}^{(n)}_{\mathcal{I}_j^{(m)}} = \bar{\Phi}_D(\{w_i^{(n)} q_i^{(n)}\}_{\mathcal{I}_j^{(m)}}), \end{aligned}$$

where $\mathcal{I}_j^{(m)}$ is a <u>sub-partition</u> of the partition $\mathcal{I}^{(m)}$ (which one has to recall to contain indices). The mixture $p^{(n)}$ is called parent mixture *of $p^{(m)}$, according to the partition $\mathcal{I}^{(m)}$.*

Note: $p^{(1)} = \bar{\Phi}_D((\boldsymbol{w}^{(n)}, \boldsymbol{q}^{(n)}))$ is the child mixture of p^n induced by the trivial partition $\mathcal{I}^{(1)} = \{[1:n]\}$. If the barycenter associativity holds, then $p^{(n)}$ and $p^{(m)}$ have the same barycenter $p^{(1)}$.

Proposition B.2.5. *Consider a D-measure and a family of distributions $\mathcal{Q}$. Then, given a mixture $p^{(n)} = (\boldsymbol{w}^{(n)})^T \boldsymbol{q}^{(n)}$ of size n, a partition $\mathcal{I}^{(m)}$ of size (m) of the interval $[1\!:\!n]$ and given $p^{(m)} = (\boldsymbol{w}^{(m)})^T \boldsymbol{q}^{(m)}$, the child mixture of*

$p^{(n)}$ induced by $\mathcal{I}^{(m)}$, the following function

$$C_D^U(p^{(n)}\|p^{(m)}) = \sum_{j=1}^{m} \sum_{i\in\mathcal{I}_j^{(m)}} w_i^{(n)} D(q_i^{(n)}\|q_j^{(m)}), \tag{B.40}$$

is an upper bound of the CTD of $p^{(m)}$ from $p^{(n)}$, i.e.

$$C_D(p^{(n)}\|p^{(m)}) \leq C_D^U(p^{(n)}\|p^{(m)}). \tag{B.41}$$

If the D-measure and family $\mathcal{Q}$ satisfy the ABTI property (3.96)*, then the following is also true*

$$C_D(p^{(n)}\|p^{(1)}) = C_D^U(p^{(n)}\|p^{(m)}) + C_D(p^{(m)}\|p^{(1)}), \tag{B.42}$$

where $p^{(1)}$ is the barycenter of both $p^{n)}$ and $p^{(m)}$ (remember that child mixtures have the same barycenter of the parents if associativity holds).

Proof. Inequality (B.41) is readily proved because the function $C_D^U(p^{(n)}\|p^{(m)})$ coincides with the function $V(W|\Theta^{(n)},\Theta^{(m)})$ (4.2) evaluated at a feasible matrix W, and by definition $C_D(p^{(n)}\|p^{(m)}) \leq V(W|\Theta^{(n)},\Theta^{(m)})$. The feasible matrix $W \in \mathcal{T}(\boldsymbol{w}^n, \boldsymbol{w}^m)$ is the following

$$W_{i,j} = w_i^n, \quad i \in \mathcal{I}_j^m, \qquad W_{i,j} = 0 \quad i \notin \mathcal{I}_j^m.$$

To prove (B.42), it is sufficient to rewrite the CTD of the barycenter $p^{(1)}$ from $p^{(n)}$ considering that $\mathcal{I}^{(m)} = \{\mathcal{I}_j^{(m)}\}_{j=1}^{(m)}$ is a partition of $[1\!:\!n]$:

$$C_D(p^{(n)}\|p^{(1)}) = \sum_{i=1}^{n} w_i^{(n)} D(q_i^{(n)}\|p^{(1)}) = \sum_{j=1}^{m} \sum_{i\in\mathcal{I}_j^{(m)}} w_i^{(n)} D(q_i^{(n)}\|p^{(1)}). \tag{B.43}$$

From the property (3.96) (ABTI)

$$\sum_{i\in\mathcal{I}_j^{(m)}} w_i^{(n)} D(q_i^{(n)}\|p^{(1)}) = \sum_{i\in\mathcal{I}_j^{(m)}} w_i^{(n)} D(q_i^{(n)}\|\hat{q}^{(n)}_{\mathcal{I}_j^{(m)}}) + \Big(\sum_{i\in\mathcal{I}_j^{(m)}} w_i^{(n)}\Big) D(\hat{q}^{(n)}_{\mathcal{I}_j^{(m)}}\|p^{(1)}), \tag{B.44}$$

and considering that, by definition of child mixture, $w_j^{(m)} = \sum_{i\in\mathcal{I}_j^{(m)}} w_i^{(n)}$ and $q_j^{(m)} = \hat{q}^{(n)}_{\mathcal{I}_j^{(m)}}$, we have

$$C_D(p^{(n)}\|p^{(1)}) = \sum_{j=1}^{m} \sum_{i\in\mathcal{I}_j^{(m)}} w_i^{(n)} D(q_i^{(n)}\|q_j^{(m)}) + \sum_{j=1}^{m} w_j^{(m)} D(q_j^{(m)}\|p^{(1)}). \tag{B.45}$$

Note that the first term of the right hand side of (B.45) is the $C_D^U(p^{(n)}\|p^{(m)})$, given by (B.40), while the second term is the $C_D(p^{(m)}\|p^{(1)})$. It follows that (B.45) coincides with (B.42). □

Remark 4. *The inequality* (B.41) *and the identity* (B.42) *in the particular case in which $m = n-1$ take a simple form. In this case, $p^{(n-1)}$ is obtained from $p^{(n)}$ by merging two (weighted) components, say $w_i^{(n)}q_i^{(n)}$ and $w_j^{(n)}q_j^{(n)}$, to obtain:*

$$C_D^U(p^{(n)}\|\tilde{p}_{i,j}^{(n)}) = w_i^{(n)}D(q_i^{(n)}\|\hat{q}_{i,j}^{(n)}) + w_j^{(n)}D(q_j^{(n)}\|\hat{q}_{i,j}^{(n)}), \tag{B.46}$$

the inequality (B.41) *takes the form*

$$C_D(p^{(n)}\|\tilde{p}_{i,j}^{(n)}) \leq w_i^{(n)}D(q_i^{(n)}\|\hat{q}_{i,j}^{(n)}) + w_j^{(n)}D(q_j^{(n)}\|\hat{q}_{i,j}^{(n)}), \tag{B.47}$$

where $\tilde{p}_{i,j}^{(n)} = p^{(n)} - (w_i^{(n)}q_i^{(n)} + w_j^{(n)}q_j^{(n)}) + (w_i^{(n)} + w_j^{(n)})\hat{q}_{i,j}^{(n)}$ is the reduced-by-one mixture obtained by merging the components i and j of $p^{(n)}$; the identity (B.42) *becomes*

$$C_D(p^{(n)}\|p^{(1)}) = w_i^{(n)}D(q_i^{(n)}\|\hat{q}_{i,j}^{(n)}) + w_j^{(n)}D(q_j^{(n)}\|\hat{q}_{i,j}^{(n)}) + C_D(\tilde{p}_{i,j}^{(n)}\|p^{(1)}), \tag{B.48}$$

Theorem B.2.6 (Minimum C_D increase merging choice)**.** *Given a mixture $p = \boldsymbol{w}^T\boldsymbol{q} \in \mathcal{Q}_{mix}$ with n components, $\{w_iq_i\}_{i=1}^n$, $n > 2$, with all weights w_i strictly positive, let $\{i^*, j^*\} \subset [1:n]$, $i^* \neq j^*$, be a pair of indexes such that*

$$\begin{aligned} w_{i^*}D(q_{i^*}\|\hat{q}_{i^*,j^*}) + w_{j^*}D(q_{j^*}\|\hat{q}_{i^*,j^*}) \leq w_iD(q_i\|\hat{q}_{i,j}) + w_jD(q_j\|\hat{q}_{i,j}),& \\ \forall\{i,j\} \subset [1:n],\ i \neq j,& \end{aligned} \tag{B.49}$$

where, as usual, $\hat{q}_{i,j} = \bar{\Phi}_D(w_iq_i + w_jq_j)$. Then, $C_D(p\|\tilde{p}_{i^*,j^*})$ and $C_D^R(p\|\tilde{p}_{i^*,j^*})$ coincide, and are as follows:

$$C_D(p\|\tilde{p}_{i^*,j^*}) = C_D^R(p\|\tilde{p}_{i^*,j^*}) = w_{i^*}D(q_{i^*}\|\hat{q}_{i^*,j^*}) + w_{j^*}D(q_{j^*}\|\hat{q}_{i^*,j^*}). \tag{B.50}$$

Moreover, any pair $\{i^*, j^*\}$ that satisfies (B.49) *also provides the least CTD increment, i.e.*

$$C_D(p\|\tilde{p}_{i^*,j^*}) \leq C_D(p\|\tilde{p}_{i,j}), \quad \forall\{i,j\} \subset [1:n]. \tag{B.51}$$

Proof. **Merging choice introducing the least C_D between contiguous model orders** The identity (B.50) is a trivial consequence of Theorem B.2.4.

It remains to prove the inequality (B.51), and this will be done by showing that the existence of a pair $\{r, s\} \neq \{i^*, j^*\}$ such that

$$C_D(p\|\tilde{p}_{rs}) < C_D(p\|\tilde{p}_{i^*,j^*}) \tag{B.52}$$

would lead to a contradiction.

Actually, it is sufficient to show that that the existence of a pair $\{r, s\} \neq \{i^*, j^*\}$ such that

$$C_D^R(p\|\tilde{p}_{r,s}) < C_D(p\|\tilde{p}_{i^*,j^*}) \tag{B.53}$$

contradicts the hypotheses, thus proving (B.51).

The computation of $C_D^R(p\|\tilde{p}_{r,s})$ can be carried out using redundant matrix representations of feasible transportation plans, i.e. matrices $W \in \mathbb{R}_+^{n\times(n+1)}$ whose columns $W_{:,r}$ and $W_{:,s}$ are set to zero. Formally, the set of redundant matrices representing feasible relaxed transportation plans is defined as follows

$$\overline{\mathcal{T}}_{r,s}(\boldsymbol{w}) = \{W \in \mathbb{R}_+^{n\times(n+1)} : \ W\mathbb{1}_{n+1} = \boldsymbol{w}, \ W_{:,r} = W_{:,s} = 0_{n\times 1}\}. \tag{B.54}$$

With this formalism

$$\begin{aligned} C_D^R(p\|\tilde{p}_{r,s}) &= \min_{W\in\overline{\mathcal{T}}_{r,s}(\boldsymbol{w})} \sum_{i=1}^{n} \left(W_{i,n+1} D(q_i\|\hat{q}_{r,s}) + \sum_{j=1}^{n} W_{i,j} D(q_i\|q_j) \right) \\ &= \sum_{i=1}^{n} \min_{W_{i,:}\mathbb{1}_{n+1}=w_i} \left(W_{i,n+1} D(q_i\|\hat{q}_{r,s}) + \sum_{\substack{j=1\\ j\neq r,s}}^{n} W_{i,j} D(q_i\|q_j) \right) \end{aligned} \tag{B.55}$$

that we can rewrite as

$$C_D^R(p\|\tilde{p}_{r,s}) = \sum_{i=1}^{n} f_{r,s}(\breve{W}_{i,:}) = \sum_{i=1}^{n} \min_{W_{i,:}\mathbb{1}_{n+1}=w_i} f_{r,s}(W_{i,:}), \tag{B.56}$$

where

$$f_{r,s}(W_{i,:}) = W_{i,n+1} D(q_i\|\hat{q}_{r,s}) + \sum_{\substack{j=1\\ j\neq r,s}}^{n} W_{i,j} D(q_i\|q_j), \quad i \in [1\!:\!n]. \tag{B.57}$$

The expression (B.56) reveals that the n rows of the optimal relaxed transportation plan $\breve{W}$ can be computed independently, by solving n independent optimization problems. It is easy to see that

$$\begin{matrix}\text{for } i \in [1\!:\!n] \\ i \neq r, s,\end{matrix} \qquad \min_{W_{i,:}\mathbb{1}_{n+1}=w_i} f_{r,s}(W_{i,:}) = 0, \text{ with } \quad \begin{cases} \breve{W}_{i,i} = w_i, & i \in [1\!:\!n], \\ \breve{W}_{i,j} = 0, & j \in [1\!:\!n+1],\ j \neq i. \end{cases} \tag{B.58}$$

Thus, the only positive contributions in the summation (B.56) that gives $C_D^R(p\|\tilde{p}_{r,s})$ are for $i = r$ and $i = s$:

$$C_D^R(p\|\tilde{p}_{r,s}) = \min_{W_{r,:}\mathbb{1}_{n+1}=w_r} f_{r,s}(W_{r,:}) + \min_{W_{s,:}\mathbb{1}_{n+1}=w_r} f_{r,s}(W_{s,:}). \tag{B.59}$$

Recalling that $C_D^R(p\|\tilde{p}_{r,s}) \leq C_D(p\|\tilde{p}_{r,s})$, and exploiting the upper bound on $C_D(p\|\tilde{p}_{r,s})$ discussed in Remark 4, we get

$$C_D^R(p\|\tilde{p}_{r,s}) \leq w_r D(q_r\|\hat{q}_{r,s}) + w_s D(q_s\|\hat{q}_{r,s}) = B_D(w_r\, q_r, w_s\, q_s). \tag{B.60}$$

Equation (B.50) tells us that for the pair $\{i^*, j^*\}$ that minimizes the bound $B_D(w_i\, q_i, w_j\, q_j)$, the bound itself coincides with $C_D^R(p\|\tilde{p}_{i^*,j^*})$, i.e.

$$C_D^R(p\|\tilde{p}_{i^*,j^*}) = w_{i^*} D(q_{i^*}\|\hat{q}_{i^*,j^*}) + w_{j^*} D(q_{j^*}\|\hat{q}_{i^*,j^*}) = B_D(w_{i^*}\, q_{i^*}, w_{j^*}\, q_{j^*}). \tag{B.61}$$

We will show that if the pair $\{r, s\}$ is not the minimizer of the bound (B.60), i.e. if $\{r, s\} \neq \{i^*, j^*\}$ and

$$w_{i^*} D(q_{i^*}\|\hat{q}_{i^*,j^*}) + w_{j^*} D(q_{j^*}\|\hat{q}_{i^*,j^*}) < w_r D(q_r\|\hat{q}_{r,s}) + w_s D(q_s\|\hat{q}_{r,s}), \tag{B.62}$$

then necessarily the following is true

$$C_D^R(p\|\tilde{p}_{i^*,j^*}) \leq C_D^R(p\|\tilde{p}_{r,s}), \tag{B.63}$$

which, recalling (B.50), implies

$$C_D(p\|\tilde{p}_{i^*,j^*}) \leq C_D(p\|\tilde{p}_{r,s}). \tag{B.64}$$

The proof proceeds by showing that the inequality

$$C_D^R(p\|\tilde{p}_{r,s}) < C_D^R(p\|\tilde{p}_{i^*,j^*}) \tag{B.65}$$

cannot be satisfied, because it would lead to a contradiction with the assumption (B.62), and therefore necessarily (B.63) must be true.

Indeed, looking at the expressions of (B.59) and (B.57), we see that the upper bound (B.60) corresponds to the following feasible transportation plan $\widetilde{W}$:

$$\widetilde{W}_{r,n+1} = w_r, \quad \widetilde{W}_{s,n+1} = w_s, \quad \text{and} \quad \begin{array}{ll} \widetilde{W}_{i,:} = \widehat{W}_{i,:}, & i \in [1\!:\!n],\ i \neq r, s, \\ \widetilde{W}_{r,j} = \widetilde{W}_{s,j} = 0, & j \in [1\!:\!n]. \end{array} \tag{B.66}$$

Indeed, there are few ways that $C_D^R(p\|\tilde{p}_{r,s})$ can be less than its bound (B.60): there may exists $h \neq s, r$, such that $D(q_r\|q_h) < D(q_r\|\hat{q}_{s,r})$, or there may exist a pair of indexes $\{h, t\} \neq \{r, s\}$ such that $D(q_r\|q_h) < D(q_r\|\hat{q}_{s,r})$ and $D(q_s\|q_t) < D(q_s\|\hat{q}_{s,r})$, in the two cases $h = t$ and $h \neq t$. We consider only one of the previous cases, because all others can be dealt with in the same way. Thus, let us consider the case where (B.65) is true because

$$\begin{aligned} \exists h \in \operatorname*{arg\,min}_{j\in[1:n]\setminus\{r,s\}} \big(D(q_r\|q_j)\big) : \quad & D(q_r\|q_h) < D(q_r\|\hat{q}_{s,r}), \\ \text{while} \quad & D(q_s\|\hat{q}_{s,r}) \leq D(q_r\|q_j), \quad \forall j \in [1:n]. \end{aligned} \tag{B.67}$$

Then, we have

$$C_D^R(p\|\tilde{p}_{r,s}) = w_r D(q_r\|q_h) + w_s D(q_s\|\hat{q}_{r,s}). \tag{B.68}$$

Note that, since the pair $\{i^*, j^*\}$ is a minimizer of the bound $B_D(w_i\, q_i, w_j\, q_j)$, considering the pair $\{h, s\}$ we have

$$w_{i^*} D(q_{i^*}\|\hat{q}_{i^*,j^*}) + w_{j^*} D(q_{j^*}\|\hat{q}_{i^*,j^*}) \leq w_h D(q_h\|\hat{q}_{h,s}) + w_s D(q_s\|\hat{q}_{h,s}). \tag{B.69}$$

On the other hand, if the inequality (B.65) were true, we would have

$$\begin{aligned} C_D^R(p\|\tilde{p}_{i^*,j^*}) =& w_r D(q_r\|q_h) + w_s D(q_s\|\hat{q}_{r,s}) < \\ <& w_{i^*} D(q_{i^*}\|\hat{q}_{i^*,j^*}) + w_{j^*} D(q_{j^*}\|\hat{q}_{i^*,j^*}) < w_r D(q_r\|\hat{q}_{r,s}) + w_s D(q_s\|\hat{q}_{r,s}). \end{aligned} \tag{B.70}$$

However, by (B.19) of Lemma B.2.2 we have

$$w_r D(q_r\|\hat{q}_{r,h}) + w_h D(q_h\|\hat{q}_{r,h}) < w_r D(q_r\|q_h) \tag{B.71}$$

Clearly, inequalities (B.70) and (B.71) are in contradiction, because put together they give

$$w_r D(q_r\|q_h) + w_s D(q_s\|\hat{q}_{r,s}) < w_r D(q_r\|q_h), \tag{B.72}$$

which is impossible because $w_s D(q_s\|\hat{q}_{r,s}) \geq 0$. Thus, (B.64) is impossible, and (B.51) is proved. □

B.2.5 Adaptive reduction theory and relative transportation loss properties

Consider sequential merging algorithms in which at each step two components are merged into one (i.e., a submixture made of two components is replaced by its barycenter). Thus, at each step of the algorithm the number of components decreases exactly by one. The sequence starts with $m = n$, at the given mixture $p^{(n)}$, and produces a sequence of *nested*[1] mixtures $p^{(m)}$, each one made of m (descending index) weighted components, denoted $\{w_i^{(m)} q_i^{(m)}\}_{i=1}^m$. At each step a submixture of two components, say $\{w_r^{(m)} q_r^{(m)}\}$ and $\{w_s^{(m)} q_s^{(m)}\}$, is suitably selected and replaced with its barycenter $\hat{q}_{r,s}^{(m)}$, with weight $w_r^{(m)} + w_s^{(m)}$, so that at a given step the mixture $p^{(m-1)}$ is obtained as

$$\begin{aligned} p^{(m-1)} = \tilde{p}_{\{r,s\}}^{(m)} &= (\boldsymbol{w}^{(m)})^T \boldsymbol{q}^{(m)} - \left(w_r^{(m)} q_r^{(m)} + w_s^{(m)} q_s^{(m)}\right) + (w_r^{(m)} + w_s^{(m)}) \hat{q}_{r,s}^{(m)} \\ &= \sum_{\substack{i=1 \\ i \neq r,s}} w_i^{(m)} q_i^{(m)} + (w_r^{(m)} + w_s^{(m)}) \hat{q}_{r,s}^{(m)}. \end{aligned} \tag{B.73}$$

Thus, a sequential merging algorithm produces a sequence of nested mixtures $p^{(m)} \in \mathcal{Q}_{\text{mix}}$, with $m = n, \dots, 1$ starting from $p^{(n)}$ and arriving at $p^{(1)}$, which is a mixture of only one component, and therefore $p^{(1)} \in \mathcal{Q}$. Note that if the associativity property (3.73) holds true, then for any choice $\{r, s\}$ of the pairs to be merged at each step, the last single component mixture $p^{(1)}$ coincides with the barycenter of the original mixture: $p^{(1)} = \bar{\Phi}_D(p^{(n)}) = \hat{p}^{(n)}$. Actually there is more: it is true that at each step $\bar{\Phi}_D(p^{(m)}) = \bar{\Phi}_D(p^{(n)})$, $m = n, \dots, 1$. If we are given the mixture $p^{(m)}$, and must choose the two components to merge, a straightforward choice is the greedy one: select the two components whose merging produces the mixture with the lowest C_D. Taking into account the notation in (B.73), the optimal local (greedy) choice of components to merge is as follows:

$$i^*, j^* \in [1\!:\!m], \quad i^* \neq j^* : \quad C_D(p^{(m)} \| \tilde{p}_{i^*,j^*}^{(m)}) \leq C_D(p^{(m)} \| \tilde{p}_{i,j}^{(m)}), \quad \forall i, j \in [1\!:\!m], \; i \neq j, \tag{B.74}$$

and setting $p^{(m-1)} = \tilde{p}_{i^*,j^*}^{(m)}$ (recall that $\tilde{p}_{i^*,j^*}^{(m)}$ is derived from $p^{(m)}$ by replacing the submixture $\{w_h^{(m)} q_h^{(m)}\}_{h \in \{i^*,j^*\}}$ with the component $\{(w_{i^*}^{(m)} + w_{j^*}^{(m)}) \hat{q}_{i^*,j^*}^{(m)}\}$.

[1] Obtained by only merging subsets of the original mixture model $p^{(n)}$.

Note that the choice of the pair (i^*, j^*) may not be unique.

This reducing strategy provides the sequence $\{p^{(m)}\}$ with the minimum increment of the CTD, i.e., at each step $C_D(p^{(m)} \| p^{(m-1)})$ is the lowest possible. For this reason, we will refer to this reduction approach as the *minimal incremental CTD strategy*, discussed in Theorem B.2.6.

Performing this iterative merging, at the m-th step each component $q_j^{(m)}$, for $j \in [1:m]$, of the mixture $p^{(m)}$ is the barycenter of a submixture of the original mixture $p^{(n)}$ defined by a set of indexes $\mathcal{I}_j^{(m)} \subset [1:n]$, where the m sets $\mathcal{I}_j^{(m)}$, define a partition $\mathcal{I}^{(m)}$ of the interval $[1:n]$:

$$\mathcal{I}^{(m)} = \{\mathcal{I}_j^{(m)}\}_{j=1}^m : \quad \bigcup_{j=1}^m \mathcal{I}_j^{(m)} = [1:n], \quad \text{and} \quad \mathcal{I}_i^{(m)} \cap \mathcal{I}_j^{(m)} = \emptyset, \quad \forall i \neq j \in [1:m]. \tag{B.75}$$

Proposition B.2.7. *For the greedy C_D binary sequential merging algorithm above presented the following holds true*

$$C_D(p^{(n)} \| p^{(1)}) = \sum_{j=m+1}^{n} C_D(p^{(j)} \| p^{(j-1)}) + C_D(p^{(m)} \| p^{(1)}). \tag{B.76}$$

Proof. Easy consequence of Proposition B.2.5 and Remark 4 □

In what follows it is useful to define the cumulative C_D increments down to the model order m:

$$s_{C_D}^{(m)} = \sum_{j=m+1}^{n} C_D(p^{(j)} \| p^{(j-1)}), \quad s_{C_D}^{(n)} = 0, \tag{B.77}$$

so that (B.76) can be rewritten as

$$C_D(p^{(n)} \| p^{(1)}) = s_{C_D}^{(m)} + C_D(p^{(m)} \| p^{(1)}). \tag{B.78}$$

Note that for the reduction according to the minimal incremental C_D criterion

$$s_{C_D}^{(1)} = C_D(p^{(n)} \| p^{(1)}). \tag{B.79}$$

Proposition B.2.8. *Given a non degenerate mixture $p^{(n)}$ with $n \geq 3$ components, the minimal C_D increment sequence $\{p^{(m)}\}_{m=n}^1$ is such that*

$$C_D(p^{(m)} \| p^{(m-1)}) \leq \frac{2}{m} C_D(p^{(m)} \| p^{(1)}), \quad m \in [n:2]. \tag{B.80}$$

where the inequality is strict for $m \in [n:3]$.

Proof. In the following $p^{(1)}$ and $\hat{q}^{(m)}$ equivalently denote the barycenter of the mixture $p^{(m)}$. Recall that the C_D of the mixture $p^{(m)}$ from $p^{(1)}$ is

$$C_D(p^{(m)} \| p^{(1)}) = \sum_{i=1}^{m} w_i^{(m)} D(q_i^{(m)} \| p^{(1)}) \tag{B.81}$$

For a given $j \in [1\!:\!m]$ this can be rewritten as

$$\begin{aligned} C_D(p^{(m)} \| p^{(1)}) = w_j^{(m)} D(q_j^{(m)} \| p^{(1)}) + & \\ + \sum_{\substack{i=1 \\ i \neq j}}^{m} \big(w_i^{(m)} D(q_i^{(m)} \| p^{(1)}) + w_j^{(m)} D(q_j^{(m)} \| p^{(1)}) - w_j^{(m)} D(q_j^{(m)} \| p^{(1)}) \big) & \\ = -(m-2) w_j^{(m)} D(q_j^{(m)} \| p^{(1)}) + \sum_{\substack{i=1 \\ i \neq j}}^{m} \big(w_i^{(m)} D(q_i^{(m)} \| p^{(1)}) + w_j^{(m)} D(q_j^{(m)} \| p^{(1)}) \big) & \end{aligned} \tag{B.82}$$

Summing for $j = 1, \ldots, m$

$$\begin{aligned} m C_D(p^{(m)} \| p^{(1)}) = -(m-2) \sum_{j=1}^{m} w_j^{(m)} D(q_j^{(m)} \| p^{(1)}) + & \\ + \sum_{j=1}^{m} \left(\sum_{\substack{i=1 \\ i \neq j}}^{m} \big(w_i^{(m)} D(q_i^{(m)} \| p^{(1)}) + w_j^{(m)} D(q_j^{(m)} \| p^{(1)}) \big) \right). & \end{aligned} \tag{B.83}$$

From the ABTI property (3.3.2), it follows:

$$\begin{aligned} w_i^{(m)} D(q_i^{(m)} \| p^{(1)}) + w_j^{(m)} D(q_j^{(m)} \| p^{(1)}) = & \\ = w_i^{(m)} D(q_i^{(m)} \| \hat{q}_{i,j}^{(m)}) + w_j^{(m)} D(q_j^{(m)} \| \hat{q}_{i,j}^{(m)}) + (w_i^{(m)} + w_j^{(m)}) D(\hat{q}_{i,j}^{(m)} \| p^{(1)}), & \end{aligned} \tag{B.84}$$

hence, equation (B.83) can be written

$$\begin{aligned} mC_D(p^{(m)} \| p^{(1)}) =& -(m-2)C_D(p^{(m)} \| p^{(1)}) \\ &+ \sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq j}}^{m} \left(w_i^{(m)} D(q_i^{(m)} \| \hat{q}_{i,j}^{(m)}) + w_j^{(m)} D(q_j^{(m)} \| \hat{q}_{i,j}^{(m)}) \right) \\ &+ \sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq j}}^{m} (w_i^{(m)} + w_j^{(m)}) D(\hat{q}_{i,j}^{(m)} \| p^{(1)}). \end{aligned} \tag{B.85}$$

Let (r, s) denote the integer pairs that satisfy the minimum C_D increment criterion, i.e.

$$\begin{aligned} w_r^{(m)} D(q_r^{(m)} \| \hat{q}_{r,s}^{(m)}) + w_s^{(m)} D(q_s^{(m)} \| \hat{q}_{r,s}^{(m)}) \leq & w_i^{(m)} D(q_i^{(m)} \| \hat{q}_{i,j}^{(m)}) + w_j^{(m)} D(q_j^{(m)} \| \hat{q}_{i,j}^{(m)}), \\ & \forall \{i, j\} \subset [1\!:\!m],\ i \neq j. \end{aligned} \tag{B.86}$$

Replacing (B.86) in (B.85) and rearranging

$$\begin{aligned} 2(m-1)C_D(p^{(m)} \| p^{(1)}) \geq & \sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq j}}^{m} \left(w_r^{(m)} D(q_r^{(m)} \| \hat{q}_{r,s}^{(m)}) + w_s^{(m)} D(q_s^{(m)} \| \hat{q}_{r,s}^{(m)}) \right) \\ &+ \sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq j}}^{m} (w_i^{(m)} + w_j^{(m)}) D(\hat{q}_{i,j}^{(m)} \| p^{(1)}). \end{aligned} \tag{B.87}$$

Recalling that $p^{(m-1)}$ is obtained from $p^{(m)}$ after merging the pair $\{w_r^{(m)} q_r^{(m)}\}$ and $\{w_s^{(m)} q_s^{(m)}\}$, and that

$$C_D(p^{(m)} \| p^{(m-1)}) = w_r^{(m)} D(q_r^{(m)} \| \hat{q}_{r,s}^{(m)}) + w_s^{(m)} D(q_s^{(m)} \| \hat{q}_{r,s}^{(m)}) \tag{B.88}$$

we get

$$\begin{aligned} 2(m-1)C_D(p^{(m)} \| p^{(1)}) \geq & m(m-1)C_D(p^{(m)} \| p^{(m-1)}) + \\ &+ \sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq j}}^{m} (w_i^{(m)} + w_j^{(m)}) D(\hat{q}_{i,j}^{(m)} \| p^{(1)}), \end{aligned} \tag{B.89}$$

and then

$$\frac{2}{m}C_D(p^{(m)}\|p^{(1)}) \geq C_D(p^{(m)}\|p^{(m-1)}) + \frac{1}{m(m-1)}\sum_{j=1}^{m}\sum_{\substack{i=1\\i\neq j}}^{m}(w_i^{(m)}+w_j^{(m)})D(\hat{q}_{i,j}^{(m)}\|p^{(1)}). \tag{B.90}$$

Since all terms in the double summations are nonnegative, and the total is strictly positive in the nonsingular case (the total is zero if and only if all components $q_i^{(m)}$ coincide), the thesis (B.80) directly follows, and the inequality is strict for $m \in [3\!:\!n]$. □

Remark 5. *Note that the inequality* (B.80) *is rather conservative whenever the mixtures* $p^{(m)}$ *are far to be singular, because the double summation in* (B.90) *in this case may be large and non negligible. Indeed, all numerical computations have shown that the increments are typically well below the upper bound* (B.80).

Proposition B.2.9. *Given a non degenerate mixture* $p^{(n)}$ *with* $n \geq 3$ *components, the cumulative* C_D *increment* $s_{C_D}^{(m)}$ *defined in* (B.77), *associated to the minimal CTD increment sequence* $\{p^{(m)}\}_{m=n}^{1}$ *is such that*

$$s_{C_D}^{(m)} \leq \rho(m,n)C_D(p^{(n)}\|p^{(1)}), \quad m \in [1\!:\!n], \tag{B.91}$$

with

$$\rho(m,n) = 1 - \frac{m(m-1)}{n(n-1)}, \quad m \in [1\!:\!n]. \tag{B.92}$$

The inequality (B.91) *is strict for* $m \in [2\!:\!n-1]$.

Proof. Consider the inequality (B.80) for $m = n$

$$C_D(p^{(n)}\|p^{(n-1)}) < \frac{2}{n}C_D(p^{(n)}\|p^{(1)}). \tag{B.93}$$

Recalling that, by definition, $s_{C_D}^{(n-1)} = C_D(p^{(n)}\|p^{(n-1)})$ and checking, from definition (B.92), that $\rho(n-1,n) = 2/n$, we have

$$s_{C_D}^{(n-1)} < \rho(n-1,n)C_D(p^{(n)}\|p^{(1)}), \tag{B.94}$$

which is (B.91) for $m = n-1$. By induction, we can prove that (B.91) holds true for all $m \in [1\!:\!n-1]$, by showing that if it holds for some $m \in [3\!:\!n-1]$, then it holds true for $m-1$. From inequality (B.80) and the definition (B.77) of the cumulative CTD increment $s_{C_D}^{(m)}$ we get

$$C_D(p^{(m)}\|p^{(m-1)}) < \frac{2}{m}\big(C_D(p^{(n)}\|p^{(1)}) - s_{C_D}^{(m)}\big), \quad m \in [3\!:\!n-1]. \tag{B.95}$$

Rearranging

$$\frac{2}{m}s_{C_D}^{(m)} + C_D(p^{(m)}\|p^{(m-1)}) < \frac{2}{m}C_D(p^{(n)}\|p^{(1)}), \quad m \in [3\!:\!n-1]. \tag{B.96}$$

Add the term $(1-2/m)s_{C_D}^{(m)}$ to both sides to get

$$s_{C_D}^{(m)} + C_D(p^{(m)}\|p^{(m-1)}) < \Big(1-\frac{2}{m}\Big)s_{C_D}^{(m)} + \frac{2}{m}C_D(p^{(n)}\|p^{(1)}), \quad m \in [3\!:\!n-1], \tag{B.97}$$

where we recognize that $s_{C_D}^{(m-1)} = s_{C_D}^{(m)} + C_D(p^{(m)}\|p^{(m-1)})$. Assuming that at step m the inequality (B.91) holds true, let us replace the term $s_{C_D}^{(m)}$ in the right-hand-side of (B.97), so that

$$\begin{aligned} s_{C_D}^{(m-1)} &< \Big(1-\frac{2}{m}\Big)\rho(m,n)C_D(p^{(n)}\|p^{(1)}) + \frac{2}{m}C_D(p^{(n)}\|p^{(1)}) \\ &= \left(\Big(1-\frac{2}{m}\Big)\rho(m,n) + \frac{2}{m}\right) C_D(p^{(n)}\|p^{(1)}), \quad m \in [3\!:\!n-1] \end{aligned} \tag{B.98}$$

To prove the theorem we only need to show that

$$\rho(m-1,n) = \Big(1-\frac{2}{m}\Big)\rho(m,n) + \frac{2}{m}, \quad m \in [3\!:\!n-1]. \tag{B.99}$$

Doing the math

$$\begin{aligned} \Big(1-\frac{2}{m}\Big)\rho(m,n) + \frac{2}{m} &= \Big(1-\frac{2}{m}\Big)\Big(1-\frac{m(m-1)}{n(n-1)}\Big) + \frac{2}{m} \\ &= 1 - \frac{2}{m} - \frac{m(m-1)}{n(n-1)} + \frac{2}{m}\frac{m(m-1)}{n(n-1)} \\ &= 1 - \frac{(m-1)(m-2)}{n(n-1)} = \rho(m-1,n), \end{aligned} \tag{B.100}$$

so that the induction is completed and the theorem is proved. □

Remark 6. *The inequality* (B.91) *has been derived on the basis of the inequality* (B.80)*, which is rather conservative, as discussed in the Remark 5. Indeed, all numerical computations have shown that the* $s_{C_D}^{(m)}$ *typically stays well below the upper bound* (B.91)*.*

B.2.6 D_{FKL} separability

D_{FKL} separability in the GGIW case

***D_{FKL}* *separability in the GGIW case*.** Let use consider two GGIW hypotheses $\zeta_i = \gamma_i \nu_i \varphi_i$ and $\zeta_j = \gamma_j \nu_j \varphi_j$. The D_{FKL} between the two components is then:

$$
\begin{aligned}
D_{FKL}(\zeta_i \| \zeta_j) &= D_{FKL}(\gamma_i \nu_i \varphi_i \| \gamma_j \nu_j \varphi_j) = \\
&= \int \gamma_i \nu_i \varphi_i \log \gamma_i \nu_i \varphi_i \mathrm{d}\chi \mathrm{d}x \mathrm{d}\mathcal{Y} - \int \gamma_i \nu_i \varphi_i \log \gamma_j \nu_j \varphi_j \mathrm{d}\chi \mathrm{d}x \mathrm{d}\mathcal{Y} = \\
&= \int \gamma_i \nu_i \varphi_i (\log \gamma_i + \log \nu_i + \log \varphi_i) \mathrm{d}\chi \mathrm{d}x \mathrm{d}\mathcal{Y} + \\
&- \int \gamma_i \nu_i \varphi_i (\log \gamma_j + \log \nu_j + \log \varphi_j) \mathrm{d}\chi \mathrm{d}x \mathrm{d}\mathcal{Y} = \\
&= \underbrace{\int \gamma_i \log \gamma_i \mathrm{d}\chi}_{-H[\gamma_i]} \underbrace{\int \nu_i \mathrm{d}x \int \varphi_i \mathrm{d}\mathcal{Y}}_{=1} + \underbrace{\int \nu_i \log \nu_i \mathrm{d}x}_{-H[\nu_i]} \underbrace{\int \gamma_i \mathrm{d}\chi \int \varphi_i \mathrm{d}\mathcal{Y}}_{=1} + \\
&+ \underbrace{\int \varphi_i \log \varphi_i \mathrm{d}\mathcal{Y}}_{-H[\varphi_i]} \underbrace{\int \nu_i \mathrm{d}x \int \gamma_i \mathrm{d}\chi}_{=1} \underbrace{- \int \gamma_i \log \gamma_j \mathrm{d}\chi}_{H_\times[\gamma_i, \gamma_j]} \underbrace{\int \nu_i \mathrm{d}x \int \varphi_i \mathrm{d}\mathcal{Y}}_{=1} + \\
&\underbrace{- \int \nu_i \log \nu_j \mathrm{d}x}_{H_\times[\nu_i, \nu_j]} \underbrace{\int \gamma_i \mathrm{d}\chi \int \varphi_i \mathrm{d}\mathcal{Y}}_{=1} \underbrace{- \int \varphi_i \log \varphi_j \mathrm{d}\mathcal{Y}}_{H_\times[\varphi_i, \varphi_j]} \underbrace{\int \nu_i \mathrm{d}x \int \gamma_i \mathrm{d}\chi}_{=1} = \\
&= \underbrace{-H[\gamma_i] + H_\times[\gamma_i, \gamma_j]}_{D_{FKL}(\gamma_i \| \gamma_j)} \underbrace{-H[\nu_i] + H_\times[\nu_i, \nu_j]}_{D_{FKL}(\nu_i \| \nu_j)} \underbrace{-H[\varphi_i] + H_\times[\varphi_i, \varphi_j]}_{D_{FKL}(\varphi_i \| \varphi_j)} = \\
&= D_{FKL}(\gamma_i \| \gamma_j) + D_{FKL}(\nu_i \| \nu_j) + D_{FKL}(\varphi_i \| \varphi_j)
\end{aligned}
\tag{B.101}
$$

□

***$B_{D_{FKL}}$* *separability in the GGIW case*.** Let us consider a GGIW intensity $p^a = (\tilde{\boldsymbol{w}}^a)^T \boldsymbol{\zeta}^a = \sum_{i=1}^{n^a} \tilde{w}_i^a \zeta_i^a$; the cost of merging two components of p^a can then be evaluated as:

$$
B_{D_{FKL}}(\tilde{w}_i^a \zeta_i^a, \tilde{w}_j^a \zeta_j^a) = \tilde{w}_i^a D_{FKL}(\zeta_i^a \| \hat{\zeta}_{i,j}^a) + \tilde{w}_j^a D_{FKL}(\zeta_j \| \hat{\zeta}_{i,j}^a), \quad \forall i, j \in [1 : n^a], \tag{B.102}
$$

where $\hat{\zeta}^a_{i,j} = \hat{\gamma}^a_{i,j}\hat{\nu}^a_{i,j}\hat{\varphi}^a_{i,j}$ is the D_{FKL}-barycenter of the pairs of components (i,j). By recalling the separability property (5.2), it follows:

$$D_{FKL}(\zeta^a_i \| \hat{\zeta}^a_{i,j}) = D_{FKL}(\gamma^a_i \| \hat{\gamma}^a_{i,j}) + D_{FKL}(\nu^a_i \| \hat{\nu}^a_{i,j}) + D_{FKL}(\varphi^a_i \| \hat{\varphi}^a_{i,j}),\ \forall i,j \in [1:n^a]. \tag{B.103}$$

By plugging now the last equation in (B.102), one obtains:

$$\begin{aligned} B_{D_{FKL}}(\tilde{w}^a_i\zeta^a_i, \tilde{w}^a_j\zeta^a_j) &= \tilde{w}^a_i\big(D_{FKL}(\gamma^a_i \| \hat{\gamma}^a_{i,j}) + D_{FKL}(\nu^a_i \| \hat{\nu}^a_{i,j}) + D_{FKL}(\varphi^a_i \| \hat{\varphi}^a_{i,j})\big) + \\ &+ \tilde{w}^a_j\big(D_{FKL}(\gamma^a_j \| \hat{\gamma}^a_{i,j}) + D_{FKL}(\nu^a_j \| \hat{\nu}^a_{i,j}) + D_{FKL}(\varphi^a_j \| \hat{\varphi}^a_{i,j})\big) = \\ &= \tilde{w}^a_i D_{FKL}(\gamma^a_i \| \hat{\gamma}^a_{i,j}) + \tilde{w}^a_j D_{FKL}(\gamma^a_j \| \hat{\gamma}^a_{i,j}) + \tilde{w}^a_i D_{FKL}(\nu^a_i \| \hat{\nu}^a_{i,j}) + \\ &+ \tilde{w}^a_j D_{FKL}(\nu^a_j \| \hat{\nu}^a_{i,j}) + \tilde{w}^a_i D_{FKL}(\varphi^a_i \| \hat{\varphi}^a_{i,j}) + \tilde{w}^a_j D_{FKL}(\varphi^a_j \| \hat{\varphi}^a_{i,j}) = \\ &= B_{D_{FKL}}(\tilde{w}^a_i\gamma^a_i, \tilde{w}^a_j\gamma^a_j) + B_{D_{FKL}}(\tilde{w}^a_i\nu^a_i, \tilde{w}^a_j\nu^a_j) + B_{D_{FKL}}(\tilde{w}^a_i\varphi^a_i, \tilde{w}^a_j\varphi^a_j) \end{aligned} \tag{B.104}$$

□

Bibliography

[1] S. Särkkä, *Bayesian Filtering and Smoothing*, ser. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2013.

[2] S. Thrun, W. Burgard, and D. Fox, *Probabilistic robotics.* MIT Press, 2005.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[4] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: Theory algorithms and software.* John Wiley & Sons, 2001.

[5] P. Billingsley, *Probability and Measure*, 2nd ed. John Wiley and Sons, 1986.

[6] L. Wasserman, *All of statistics : a concise course in statistical inference.* Springer, 2010.

[7] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[8] R. P. S. Mahler, *Statistical Multisource-Multitarget Information Fusion.* Artech House, Inc., 2007.

[9] 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)*, vol. 39, no. 1, pp. 1–38, 1977.

[10] S. Chander and P. Vijaya, “Unsupervised learning methods for data clustering,” in *Artificial Intelligence in Data Mining*, D. Binu and B. Rajakumar, Eds. Academic Press, 2021, pp. 41–64.

[11] Y. Bar-Shalom, P. Willett, and X. Tian, *Tracking and Data Fusion: A Handbook of Algorithms.* YBS Publishing, 2011.

[12] D. Alspach and H. Sorenson, “Nonlinear Bayesian estimation using Gaussian sum approximations,” *IEEE Transactions on Automatic Control*, vol. 17, no. 4, pp. 439–448, 1972.

[13] S. Blackman, “Multiple hypothesis tracking for multiple target tracking,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 19, no. 1, pp. 5–18, Jan. 2004.

[14] Z. Robotka and A. Zempléni, “Image retrieval using Gaussian mixture models,” *Annales Univ. Sci. Budapest., Sect. Comp*, vol. 31, pp. 93–105, 01 2009.

[15] H. Chui and A. Rangarajan, “A feature registration framework using mixture models,” in *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. MMBIA-2000 (Cat. No.PR00737)*, 2000, pp. 190–197.

[16] K. P. Murphy, *Probabilistic Machine Learning: An introduction.* MIT Press, 2022.

[17] R. O. Duda and P. E. Hart, “Pattern classification and scene analysis,” in *A Wiley-Interscience publication*, 1973.

[18] T. Ardeshiri, E. Ozkan, and U. Orguner, “On reduction of mixtures of the exponential family distributions,” 2013.

[19] A. D’Ortenzio, C. Manes, and U. Orguner, “An optimal transport perspective on gamma Gaussian inverse-Wishart mixture reduction,” in *IEEE 25th International Conference on Information Fusion (FUSION 2022)*, Linköping, Sweden, July 2022, pp. 1–8.

[20] K. Granström, A. Natale, P. Braca, G. Ludeno, and F. Serafino, “Gamma Gaussian inverse Wishart probability hypothesis density for extended target tracking using X-band marine radar data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6617–6631, Dec. 2015.

[21] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[22] 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.

[23] H. Sorenson and D. Alspach, “Recursive Bayesian estimation using Gaussian sums,” *Automatica*, vol. 7, no. 4, pp. 465–479, 1971.

[24] C. Lundquist, K. Granström, and U. Orguner, “An extended target CPHD filter and a gamma Gaussian inverse Wishart implementation,” *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 3, pp. 472–483, Jun. 2013.

[25] B.-N. Vo and W.-K. Ma, “The Gaussian mixture probability hypothesis density filter,” *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4091–4104, 2006.

[26] K. Granström, M. Fatemi, and L. Svensson, “Poisson multi-Bernoulli mixture conjugate prior for multiple extended target filtering,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 7, no. 1, pp. 208–225, Feb. 2020.

[27] K. Granström and U. Orguner, “Estimation and maintenance of measurement rates for multiple extended target tracking,” in *IEEE 15th International Conference on Information Fusion (FUSION)*, 2012, pp. 2170–2176.

[28] D. R. Hunter and K. L. Lange, “A tutorial on MM algorithms,” *The American Statistician*, vol. 58, pp. 30 – 37, 2004.

[29] S. Lloyd, “Least squares quantization in pcm,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[30] S. J. Julier and J. K. Uhlmann, “New extension of the Kalman filter to nonlinear systems,” in *Defense, Security, and Sensing*, 1997.

[31] I. Arasaratnam and S. Haykin, “Cubature Kalman filters,” *IEEE Transactions on Automatic Control*, vol. 54, no. 6, pp. 1254–1269, 2009.

[32] N. J. Gordon, D. Salmond, and A. F. M. Smith, “Novel approach to nonlinear/non-Gaussian Bayesian state estimation,” 1993.

[33] A. F. García-Fernández, J. L. Williams, K. Granström, and L. Svensson, "Poisson multi-bernoulli mixture filter: Direct derivation and implementation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 4, pp. 1883–1901, 2018.

[34] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Ann. of Math. Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[35] A. D'Ortenzio and C. Manes, "Likeness-based dissimilarity measures for Gaussian mixture reduction and data fusion," in *IEEE 24th International Conference on Information Fusion (FUSION 2021)*, Sun City, South Africa, November 2021, pp. 1–8.

[36] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives.* Springer Publishing Company, Incorporated, 2010.

[37] J. Williams and P. Maybeck, "Cost-function-based Gaussian mixture reduction for target tracking," in *Sixth International Conference of Information Fusion, 2003. Proceedings of the*, vol. 2, 2003, pp. 1047–1054.

[38] M. Liu, B. C. Vemuri, S.-I. Amari, and F. Nielsen, "Shape retrieval using hierarchical total Bregman soft clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2407–2419, 2012.

[39] K. Kampa, E. Hasanbelliu, and J. C. Principe, "Closed-form Cauchy-Schwarz pdf divergence for mixture of Gaussians," in *The 2011 International Joint Conference on Neural Networks*, 2011, pp. 2578–2585.

[40] F. Nielsen, "Closed-form information-theoretic divergences for statistical mixtures," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, pp. 1723–1726.

[41] D. J. Petrucci, "Gaussian mixture reduction for Bayesian target tracking in clutter," Master's thesis, Air Force Inst. of Tech., 2005.

[42] A. Ben Hamza and H. Krim, "Jensen-Renyi divergence measure: theoretical and computational perspectives," in *IEEE International Symposium on Information Theory, 2003. Proceedings.*, 2003, pp. 257–257.

[43] T. Van Erven and P. Harremos, “Rényi divergence and Kullback-Leibler divergence,” *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[44] F. Wang, T. Syeda-Mahmood, B. C. Vemuri, D. Beymer, and A. Rangarajan, “Closed-form Jensen-Renyi divergence for mixture of Gaussians and applications to group-wise shape registration,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, Eds. Springer Berlin Heidelberg, 2009, pp. 648–655.

[45] B. Fuglede and F. Topsoe, “Jensen-Shannon divergence and Hilbert space embedding,” in *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, 2004, pp. 31–.

[46] A. D’Ortenzio and C. Manes, “Consistency issues in Gaussian mixture models reduction algorithms,” 2021. [Online]. Available: https://arxiv.org/abs/2104.12586

[47] Y. Xu, Y. Fang, W. Peng, and Y. Wu, “An efficient Gaussian sum filter based on prune-cluster-merge scheme,” *IEEE Access*, vol. 7, pp. 150 992–151 005, 2019.

[48] A. D’Ortenzio, C. Manes, and U. Orguner, “Fixed-point iterations for several dissimilarity measure barycenters in the Gaussian case,” 2022. [Online]. Available: https://arxiv.org/abs/2205.04806

[49] T. Minka, “Divergence measures and message passing,” Microsoft, Tech. Rep. MSR-TR-2005-173, January 2005.

[50] A. Runnalls, “Kullback-Leibler approach to Gaussian mixture reduction,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 3, pp. 989–999, Jul. 2007.

[51] F. Nielsen and R. Nock, “Sided and symmetrized Bregman centroids,” *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2882–2904, Jun. 2009.

[52] G. Battistelli and L. Chisci, “Kullback–Leibler average, consensus on probability densities, and distributed state estimation with guaranteed stability,” *Automatica*, vol. 50, no. 3, pp. 707–718, 2014.

[53] P. C. Álvarez Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán, "A fixed-point approach to barycenters in Wasserstein space," *Journal of Mathematical Analysis and Applications*, vol. 441, no. 2, pp. 744–762, 2016.

[54] A. Assa and K. N. Plataniotis, "Wasserstein-distance-based Gaussian mixture reduction," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1465–1469, 2018.

[55] Y. Chen, T. T. Georgiou, and A. Tannenbaum, "Optimal transport for Gaussian mixture models," *IEEE Access*, vol. 7, pp. 6269–6278, 2019.

[56] D. J. Salmond, "Mixture reduction algorithms for target tracking in clutter," in *Signal and Data Processing of Small Targets 1990*, O. E. Drummond, Ed., vol. 1305, International Society for Optics and Photonics. SPIE, 1990.

[57] M. A. West, "Approximating posterior distributions by mixtures," *Journal of the royal statistical society series b-methodological*, vol. 55, pp. 409–422, 1993.

[58] J. L. Williams and P. S. Maybeck, "Cost-function-based hypothesis control techniques for multiple hypothesis tracking," *Mathematical and Computer Modelling*, vol. 43, no. 9, pp. 976–989, May 2006.

[59] H. D. Chen, K. C. Chang, and C. Smith, "Constraint optimized weight adaptation for Gaussian mixture reduction," in *Signal Processing, Sensor Fusion, and Target Recognition XIX*, I. Kadar, Ed., vol. 7697, International Society for Optics and Photonics. SPIE, 2010, pp. 281 – 290.

[60] D. F. Crouse, P. Willett, K. Pattipati, and L. Svensson, "A look at Gaussian mixture reduction algorithms," in *14th International Conference on Information Fusion*, 2011, pp. 1–8.

[61] T. Ardeshiri, U. Orguner, and E. Özkan, "Gaussian mixture reduction using reverse Kullback-Leibler divergence," 2015. [Online]. Available: https://arxiv.org/abs/1508.05514

[62] A. D'Ortenzio and C. Manes, "Composite transportation dissimilarity in consistent Gaussian mixture reduction," in *IEEE 24th International*

Conference on Information Fusion (FUSION 2021), Sun City, South Africa, November 2021, pp. 1–8.

[63] D. W. Scott, "Parametric statistical modeling by minimum integrated square error," *Technometrics*, vol. 43, no. 3, pp. 274–285, 2001.

[64] D. W. Scott and W. F. Szewczyk, "From kernels to mixtures," *Technometrics*, vol. 43, no. 3, pp. 323–335, 2001.

[65] L. Pishdad and F. Labeau, "A new reduction scheme for Gaussian sum filters," *2014 48th Asilomar Conference on Signals, Systems and Computers*, pp. 1351–1357, 2014.

[66] N. Vasconcelos and A. Lippman, "Learning mixture hierarchies," in *Advances in Neural Information Processing Systems*, M. Kearns, S. Solla, and D. Cohn, Eds., vol. 11. MIT Press, 1998.

[67] J. Goldberger and S. Roweis, "Hierarchical clustering of a mixture model," in *Advances in Neural Information Processing Systems*, L. Saul, Y. Weiss, and L. Bottou, Eds., vol. 17. MIT Press, 2004.

[68] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, no. 58, pp. 1705–1749, 2005.

[69] K. Zhang and J. Kwok, "Simplifying mixture models through function approximation," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19. MIT Press, 2006.

[70] D. Schieferdecker and M. F. Huber, "Gaussian mixture reduction via clustering," in *2009 12th International Conference on Information Fusion*, 2009, pp. 1536–1543.

[71] Y. Bar-Yosef and Y. Bistritz, "Gaussian mixture models reduction by variational maximum mutual information," *IEEE Transactions on Signal Processing*, vol. 63, no. 6, pp. 1557–1569, 2015.

[72] M. A. Brubaker, A. Geiger, and R. Urtasun, "Map-based probabilistic visual self-localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 4, pp. 652–665, 2016.

[73] J. R. Hershey and P. A. Olsen, “Approximating the Kullback-Leibler divergence between Gaussian mixture models,” in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, 2007, pp. IV–317–IV–320.

[74] L. Yu, T. Yang, and A. B. Chan, “Density-preserving hierarchical EM algorithm: Simplifying Gaussian mixture models for approximate inference,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pp. 1323–1337, 2019.

[75] M. F. Huber and U. D. Hanebeck, “Progressive Gaussian mixture reduction,” in *2008 11th International Conference on Information Fusion*, 2008, pp. 1–8.

[76] P. Bruneau, M. Gelgon, and F. Picarougne, “Parsimonious reduction of Gaussian mixture models with a variational-Bayes approach,” *Pattern Recognition*, vol. 43, no. 3, pp. 850–858, 2010.

[77] L. Ambrosio and N. Gigli, *A User's Guide to Optimal Transport.* Springer Berlin Heidelberg, 2013, pp. 1–155.

[78] J. Pitrik and D. Virosztek, “On the joint convexity of the Bregman divergence of matrices,” *Letters in Mathematical Physics*, vol. 105, pp. 675–692, 2015.

[79] P. Moulin and V. V. Veeravalli, *Statistical Inference for Engineers and Data Scientists.* Cambridge University Press, 2018.

[80] A. D'Ortenzio, C. Manes, and U. Orguner, “A model selection criterion for the mixture reduction problem based on the Kullback-Leibler divergence,” in *IEEE 25th International Conference on Information Fusion (FUSION 2022)*, Linköping, Sweden, July 2022, pp. 1–8, Winner of the Tammy L. Blair Best Student Paper Award.

[81] J. W. Koch, “Bayesian approach to extended object and cluster tracking using random matrices,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 44, no. 3, pp. 1042–1059, Jul. 2008.

[82] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems.* Norwood, MA: Artech House, 1999.

[83] D. Reid, “An algorithm for tracking multiple targets,” *IEEE Trans. Autom. Control*, vol. 24, no. 6, pp. 843–854, Dec. 1979.

[84] K. Granström, M. Baum, and S. Reuter, “Extended object tracking: Introduction, overview, and applications,” *Journal of Advances in Information Fusion*, vol. 12, no. 2, pp. 139–174, Dec. 2017.

[85] K. Granström and U. Orguner, “On the reduction of Gaussian inverse Wishart mixtures,” in *IEEE 15th International Conference on Information Fusion (FUSION)*, 2012, pp. 2162–2169.

[86] M. Rosenblatt, “Remarks on Some Nonparametric Estimates of a Density Function,” *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832 – 837, 1956.

[87] E. Parzen, “On Estimation of a Probability Density Function and Mode,” *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065 – 1076, 1962.

[88] F. Nielsen and S. Boltz, “The Burbea-Rao and Bhattacharyya centroids,” *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5455–5466, 2011.