



UNIVERSITÉ CATHOLIQUE DE LOUVAIN
FACULTÉ DES SCIENCES APPLIQUÉES
DÉPARTEMENT D'ÉLECTRICITÉ

MICROELECTRONICS LAB. AND MACHINE LEARNING GROUP

Contrast properties of entropic criteria for Blind Source Separation

**A unifying framework based on information-theoretic
inequalities**

Frédéric D. Vrins

Thesis submitted in partial fulfillment
of the requirements for the degree of
Docteur en sciences appliquées

Dissertation committee:

Prof. Michel Verleysen (Microelectronics Laboratory, UCL), advisor
Prof. Philippe Delsarte (Department of Informatics, UCL)
Prof. Christian Jutten (Institut National Polytechnique de Grenoble)
Prof. Philippe Lambert (Université de Liège)
Prof. Erkki Oja (Helsinki University of Technology)
Prof. Jean-Didier Legat (Dean of the Faculty, UCL), President

CONTRAST PROPERTIES OF ENTROPIC CRITERIA FOR BLIND SOURCE SEPARATION

A unifying framework based on information-theoretic inequalities

Frédéric D. Vrins

Faculté des Sciences Appliquées
Département d'Électricité
Microelectronics Lab. and Machine Learning Group



**Université catholique de Louvain
Louvain-la-Neuve (Belgium)**

*To my grand-fathers,
my plum and my son*

CONTENTS

Abstract	xiii
Acknowledgments	xv
Acronyms	xvii
List of Notation	xix
Introduction	xxiii
1 BSS and its relationship to ICA	1
1.1 BSS: Motivation	2
1.2 ICA: an efficient tool for BSS	7
1.2.1 PD-equivelency and Non-mixing matrices	7
1.2.2 Independence and ICA	11
1.2.3 ICA and BSS	12
1.3 Independence measures	14
1.3.1 Divergence measures between densities	14
1.3.1.1 KL properties	15
1.3.1.2 From KL to mutual information	16
1.3.2 Other measures of independence	17
1.4 Extraction schemes and contrast function definition	18
1.4.1 Extraction schemes	19
1.4.1.1 Simultaneous separation	19
1.4.1.2 Deflation separation	19
1.4.1.3 Partial separation	19
1.4.2 Contrast functions	20
1.4.2.1 Simultaneous separation	20
1.4.2.2 Deflation separation	20
1.4.2.3 Partial separation	21
1.5 Whitening preprocessing and geodesic search	22
1.5.1 Whitening	22

1.5.2	Orthogonal contrast functions	24
1.5.3	Angular parametrization in the K=2 case	25
1.5.4	Manifold-constrained problem and geodesic optimization	25
1.6	Adaptive maximization of contrast functions	27
1.7	BSS and information measures	29
1.7.1	Information measure	30
1.7.1.1	Coding using Hartley's formula	30
1.7.1.2	Information and entropy	32
1.7.1.3	Extension to continuous random variables	33
1.7.1.4	Information gain and Mutual information	34
1.7.2	Entropy as a “complexity measure”	37
1.7.3	Generalized information measures	40
1.8	Issues and objectives of the Thesis	42
2	Contrast property of Entropic criteria	45
2.1	Some tools for building contrast functions	47
2.1.1	From orthogonal deflation to orthogonal partial separation	47
2.1.2	Huber's superadditivity	49
2.2	Shannon's entropy-based contrast	51
2.2.1	Simultaneous approach	51
2.2.2	Deflation approach	53
2.2.2.1	The contrast property	53
2.2.2.2	Non-mixing local maxima	57
2.2.3	Partial approach	58
2.3	Minimum range contrast	59
2.3.1	Support and Brunn-Minkowski Inequality	59
2.3.2	Properties of the range	60
2.3.3	Simultaneous approach	61
2.3.4	Deflation approach	62
2.3.5	Partial approach	64
2.3.6	Support versus Range	65
2.3.7	A tool for building a D-BSS contrast based on Huber	65
2.4	Rényi's entropy contrast	66
2.4.1	Taylor development of Rényi's entropy	67
2.4.2	Deflation approach	70
2.4.3	Simultaneous approach	71
2.4.4	Partial approach	72
2.4.5	Some counter-examples	72

2.5 Conclusion of the chapter	78
2.5.1 Summary of results	78
2.5.2 Use of Rényi entropies in blind separation/deconvolution	79
2.6 Appendix: Proofs of results of the Chapter	84
2.6.1 Proof of Corollary 6 (wording p. 50)	84
2.6.2 Proof of Property 4 (wording p. 57)	84
2.6.3 Proof of Theorem 11 (wording p. 58)	86
2.6.4 Proof of Lemma 6 (wording p. 60)	87
2.6.5 Proof of Theorem 14 (wording p. 62)	89
2.6.6 Proof of Theorem 15 (wording p. 63)	90
2.6.7 Proof of Theorem 16 (wording p. 64)	91
2.6.8 Convolution of Gaussian kernels (wording p. 67)	92
2.6.9 Proof of Lemma 7 (wording p. 71)	93
2.6.10 Proof of Lemma 8 (wording p. 71)	94
2.6.11 Proof of Lemma 9 (wording p. 72)	95
2.6.12 Proof of Lemma 10 (wording p. 72)	95
3 Discriminacy property of Entropic contrasts	97
3.1 Concept definition and justification	99
3.2 Discriminacy of Shannon's entropy	100
3.2.1 Informal approach : the multimodal case	102
3.2.1.1 Location of the entropy minima	102
3.2.1.2 Modality and entropy minima	107
3.2.2 Formal analysis using a Taylor expansion	112
3.2.2.1 Simultaneous (mutual information)	112
3.2.2.2 Deflation (negentropy)	119
3.2.3 Formal analysis using entropy approximation	120
3.2.3.1 Entropy bounds on multimodal pdf	121
3.2.3.2 Mixing local minima in multimodal BSS	125
3.2.3.3 Local minimum points of $H(\mathbf{w}U)$	126
3.2.3.4 Local minimum points of $h(\mathbf{w}S)$	129
3.2.3.5 Complementary observations	133
3.2.4 Cumulant-based versus Information-theoretic approaches	133
3.3 Discriminacy of Rényi's entropy	138
3.4 Discriminacy of the minimum range approach	139
3.4.1 Preliminaries : $K = 2$ case	139
3.4.2 Deflation approach	140
3.4.2.1 Small variation analysis on manifolds	140

3.4.2.2	Piecewise g-convexity on hyper-sphere	143
3.4.2.3	Second order approach	149
3.4.3	Simultaneous approach	150
3.4.4	Partial approach	152
3.4.5	Givens trajectories and discriminacy	155
3.5	Discriminacy of the minimum support approach	160
3.6	Summary of results and contrast sets configuration	163
3.7	Conclusion of the chapter	164
3.7.1	Summary of results	164
3.7.2	Comparison with existing results	166
3.8	Appendix: Proofs of results of the Chapter	167
3.8.1	Proof of relation (3.14) (wording p. 114)	167
3.8.2	Proof of Lemma 11 (wording p. 116)	169
3.8.3	Proof of Lemma 12 (wording p. 116)	170
3.8.4	Proof of Lemma 13 (wording p. 121)	171
3.8.5	Proof of Lemma 14 (wording p. 126)	173
3.8.6	Proof of Lemma 15 (wording p. 130)	174
3.8.7	Proof of Lemma 16 (wording p. 140)	177
3.8.8	Proof of Theorem 18 (wording p. 141)	178
3.8.9	Proof of Lemma 17 (wording p. 149)	179
3.8.10	Proof of Lemma 19 (wording p. 151)	180
3.8.11	Proof of Corollary 15 (wording p. 152)	181
3.8.12	Proof of Lemma 20 (wording p. 152)	181
3.8.13	Proof of Lemma 21 (wording p. 152)	181
3.8.14	Proof of Lemma 23 (wording p. 154)	185
3.8.15	Proof of Corollary 18 (wording p. 155)	186
4	Minimum output range methods	187
4.1	Minimum range approach	188
4.1.1	Interpretation of the simultaneous approach	189
4.1.1.1	Interpretation in the mixture space	189
4.1.1.2	Interpretation in the output space	190
4.1.2	Interpretation of the deflation approach	191
4.2	Range estimation	198
4.2.1	Some existing methods for endpoint estimation	198
4.2.2	Existing Range estimation	200
4.2.2.1	Support estimation via density estimation	202
4.2.2.2	Range estimation for BSS application	203

4.2.3	Quasi-range based approach	203
4.2.3.1	The observed range estimator	204
4.2.3.2	The m-averaged quasi-range estimator	208
4.2.3.3	Robustness of minimum-support ICA algorithms	213
4.3	Range minimization algorithm: SWICA	220
4.3.1	Algorithm	222
4.3.2	Performance analysis of SWICA for OS-based range estimators	224
4.4	Extensions of the minimum range	226
4.4.1	The problem of blind images separation : NOSWICA	227
4.4.1.1	SWICA for correlated images separation	227
4.4.1.2	NOSWICA: a non-orthogonal extension	232
4.4.2	Application to lower- or upper-bounded sources with possible infinite range	238
4.4.2.1	LABICA	238
4.4.2.2	Practical estimation	241
4.4.2.3	Optimization scheme for “hard” ICA problems	242
4.4.2.4	LABICA applied to MLSP’06 benchmark	246
4.5	Conclusion of the Chapter	250
4.6	Appendix	250
4.6.1	Proof of relation (4.52)	250
4.6.2	Expectation of the order statistics cdf differences	251
4.6.3	Variance of the order statistics cdf differences	253
5	Conclusion	255
Appendix A:	Announcement of the IEEE MLSP 2006 data analysis competition	275
Appendix B:	Author’s publication list	279

ABSTRACT

In the recent years, Independent Component Analysis (ICA) has become a fundamental tool in adaptive signal and data processing, especially in the field of Blind Source Separation (BSS). Even though there exist some methods for which an algebraic solution to the ICA problem may be found, other iterative methods are very popular. Among them is the class of information-theoretic approaches, laying on entropies. The associated objective functions are maximized based on optimization schemes, and on gradient-ascent techniques in particular. Two major issues in this field are the following: 1) Does the global maximum point of these entropic objectives correspond to a satisfactory solution of BSS ? and 2) as gradient techniques are used, optimization algorithms look in fact for local maximum points, so what about the meaning of these local optima from the BSS problem point of view?

Even though there are some partial answers to these questions in the literature, most of them are based on simulations and conjectures; formal developments are often lacking. This thesis aims at filling this lack as well as providing intuitive justifications, too. We focus the analysis on Rényi's entropy-based contrast functions. Our results show that, generally speaking, Rényi's entropy is not a suitable contrast function for BSS, even though we recover the well-known results saying that Shannon's entropy-based objectives are contrast functions. We also show that the range-based contrast functions can be built under some conditions on the sources.

The BSS problem is stated in the first chapter, and viewed under the information (theory) angle. The two next chapters address specifically the above questions. Finally, the last chapter deals with range-based ICA, the only “entropy-based contrast” which, based on the enclosed results, is also a *discriminant* contrast function, in the sense that it is theoretically free of spurious local optima. Geometrical interpretations and surprising examples are given. The interest of this approach is confirmed by testing the algorithm on the MLSP 2006 data analysis competition benchmark; the proposed method outperforms the previously obtained results on large-scale and noisy mixture samples obtained through ill-conditioned mixing matrices.

ACKNOWLEDGMENTS

Do you know a PhD advisor who reads your paper drafts until 3 AM ? Do you know a PhD advisor who corrects the typos in your thesis on Sunday PM ? Do you know a PhD advisor who invites its whole research team for a BBQ in his garden on a sunny Sunday afternoon ? Probably not, because the subset formed by such advisors is very small and spread over the set of worldwide higher degree professors. Prof. Michel Verleysen belongs to this class. In addition to his scientific and teaching skills, he has also, as you now guess, many human skills. Even regarding its administrative duties in the Applied Sciences Faculty, it was a pleasure to collaborate with him. Thank you, dear Michel, for these wonderful four years that we have shared to work, discuss, and joke. I hope that they were as good for you than they were for me.

I take the opportunity to thank other famous professors. First, Prof. Christian Jutten, from the INPG. Our discussions and his pertinent remarks and advises along my thesis have certainly contributed to the quality of this work. In particular, I would like to warmly thank him, as well as his wife Brigitte, for their friendly welcome when I joined the INPG for two weeks in Fall 2005. All my gratitude also goes to Prof. Dinh-Tuan Pham from the INPG/CNRS. My collaboration with him dates back to the fifth ICA conference in Granada (Spain), where we had discussed on a way to rigorously prove the conjectured existence of spurious optima in ICA methods. From this date to now, I have had the pleasure and the honor to work with this outstanding statistician, and to visit him at Grenoble (where I was given an office with unforgettable sunshine sights on the Vercors mountain). I've learned a lot on his sides, specifically in mathematics and statistics. This joint work led us to co-sign several (of my best) papers: a large part of the material of chapters 2 and 3 results from our fruitful collaboration. Thank you, Tuan.

I would like to thank Prof. Ph. Delsarte, Prof. Ph. Lambert and Prof. E. Oja (an author of the most popular book on ICA) for having accepted to join Michel and Christian in my thesis jury, as well as the Dean of the Applied Science Faculty, Jean-Didier Legat, for having accepted to head them. I mention in passing that it was my pleasure to collaborate with him when I was president of the ACSSA (researchers association).

I thank my (former and present) colleagues of the UCL Machine Learning Group, and more specifically: Nicolas Donckers, John Lee and Cérdic Archambeau. They are also really good friends, with who we can rethink the world around a CMOS transistor, a *Maxi Chicken* Burger-Orval-Picon, cranberry juice or a bottle of whisky, depending on the guy. I also acknowledge Yannick (from *Les Marçels*), François and Eric for our “strolling discussions” at lunchtime, as well as all my friends, in particular the BTCZ power (oss). I do not forget to acknowledge Jean-Pierre Palamin from the Collège Cardinal Mercier; he was my favorite physics teacher at high school. I remember our experiment about the “tire-bouchon droitier” rule; it was probably the Big Bang who decides me to go further in the field of applied sciences.

This thesis wouldn’t exist without my family. By fits and starts, the PhD thesis was the only way (acceptable for me) to also receive the “Doctor” degree without having to cut human in pieces, just like my grand father, my father, my godfather, my uncle and my sister have done, do or will do. More concretely, I am really grateful to my parents for having helped me to support the printing fees of this work. Special thanks to my godfather for discussion about the world of scientific research and fetal health monitoring. All my friendship also goes to my family-in-law, for gripping discussions we have about sciences and other stuffs.

Last but not least, I would like to thank Emmanuelle, my wife, for her encouragements and her love. Sorry for having forgotten some days to let, as you said so well, my balluchon of unsolved equations in my office. Thank you my darling for having given to us our wonderful little monster which constitutes, with you, the core of my life. This thesis is also obviously dedicated to both of you.

Other special thanks to R.M. Gray (Stanford University) for discussion on self-information, E. Lutwak (Brooklyn University) for discussion on Hölder’s inequality, C. Arndt (Ford inc.) for discussion on the concept of “information”), J.-F. Cardoso (Ecole Normale Supérieure des Télécommunications) for discussion on ICA and multimodal densities, D. Erdogmus (Oregon Health & Science inst.) for discussion on m-spacings, M. Saerens (UCL) for having provided useful references, A. Meister (Australian National University) for having kindly communicated the preprint [Meister, 2006] during the review process, P.A. Absil (UCL) for discussion on manifolds and for having provided the preprint [Absil et al.], J.A. Lee (UCL) for having collaborate on optimization schemes of some contrast functions as well as all of those I’ve forgotten at the time I was writing these lines.

F.D.V.

ACRONYMS

BMI	Brunn-Minkowski Inequality
BSS	Blind Source Separation
CDF	Cumulative Density Function
D-BSS	Deflation BSS
DS	Darmois-Skitovitch
EPI	Entropy Power Inequality
ERE	Extended Renyi's Entropy
GJS	Generalized Jensen-Shannon
ICA	Independent Component Analysis
IT	Information Theory (or information-theoretic)
KL	Kullback-Leibler divergence
MI	Mutual Information
OS	Order Statistics
P-BSS	Partial BSS
PCA	Principal Component Analysis
PDF	Probability Density Function
RV	Random Variable
S-BSS	Simultaneous BSS
SIR	Signal-to-Interference ratio
SNR	Signal-to-Noise ratio
SPI	Square Performance Index

LIST OF NOTATION

Typesetting convention

m, n, k, K	Integers (lower/upper case letter)	5
α, β	Strictly positive scalar (lower case Greek letter)	12
X, x	Random variable (uppercase) and one of its sample (lowercase)	5
\mathbf{X}	Vector of random variables (uppercase, boldface)	3
X_i	i -th component of \mathbf{X} (uppercase with index in subscript)	3
$X_{(i:N)}$	i -th order statistic of X from N samples	199
$x_{(i:N)}$	i -th largest value of X from a sample set of size N	198
\mathbf{w}	Row vector (lower case, boldface)	19
$\mathbf{w}(i) = w_i$	i -th entry of vector \mathbf{w}	30
\mathbf{M}	Matrix (uppercase,boldface)	4
$\mathbf{m}_i = [\mathbf{M}]_i$	Vector, i -th row of matrix \mathbf{M} (lowercase, boldface with index in subscript)	19
$M_{ij} = [\mathbf{M}]_{ij}$	General (i, j) element of matrix \mathbf{M}	18

Statistical operators

$\text{Ex}[.]$	Expectation operator with respect to the PDF of X	5
$\text{Var}[.]$	Variance operator	11
$\text{Corr}[., .]$	Correlation operator	11
$\text{Cov}[., .]$	Covariance operator	6

Particular functions

$\mu[.]$	Lebesgue measure of set	41
$\Pr(.)$	Probability measure	31
$\Pr_X(.)$	CDF of X	54

$p_X(\cdot)$	PDF of X	6
$\phi(\cdot)$	Standardized univariate Gaussian function	55
$\Phi(\cdot)$	Cumulative distribution function of ϕ	171
$\phi_X(\cdot)$	Univariate Gaussian density function with zero-mean and bandwidth equal to $\sqrt{\text{Var}[X]}$	6
$\phi_{\mathbf{X}}(\cdot)$	Multivariate Gaussian density function with zero-mean and same covariance matrix as X	34
$\psi_X(x)$	Score function of X	57
$\psi_r(X)$	r -score function of X ($\psi_{r,X}(x)$: $\psi_r(X)$ at $X = x$)	69
e	$\exp(1)$	7
$\text{Erf}(\cdot)$	Error function	122
$\text{Erfc}(\cdot)$	Complementary error function	122

Matrix and vectors operations

$\text{rank}(\cdot)$	Rank operator	6
$\ \mathbf{w}\ $	Euclidean norm (for vector argument); Frobenius norm (for matrix argument)	38
T	Transpose	3
\sim	PD-equivelency operator between two matrices	8
\sim_u	SubPD-equivelency symbol	10

Particular matrices, vectors and sets

A	Mixing matrix	4
V	Whitening matrix	22
B	Demixing matrix	4
W	Global (transfer) matrix	18
Λ	Diagonal matrix	8
Π	Permutation matrix	7
P	Projection matrix	145
I_K	$K \times K$ identity matrix	7
\mathbf{e}_i	shorthand notation for $[\mathbf{I}_K]_i$	8
$\hat{\mathbf{w}}$	subvector of $\mathbf{w} \in \mathbb{R}^K$ with respect to a set of index I , $1 \leq \#I \leq K$	147
$\mathbf{1}_{\pm}^{k,n}$	k-entries vector with elements in $\{-1, 1\}$ and $\pm \mathbf{1}^{k,n} = \pm \mathbf{1}^{k,m}$ iff $m = n$	146

$\mathcal{S}(K)$	Set of K -entries unit-norm vectors	38
$\mathcal{M}(K)$	General Linear Group (group of $K \times K$ regular matrices) and its associated subspace in $\mathbb{R}^{K \times K}$	5
$\mathcal{D}(K)$	Diagonal Group (group of $K \times K$ diagonal matrices)	8
$\mathcal{P}(K)$	Permutation Group (group of $K \times K$ permutation matrices) ..	8
$\mathcal{W}(K)$	Set of $K \times K$ non-mixing matrices	7
$\mathcal{W}^{P \times K}$	Set of $P \times K$ non-mixing matrices	7
$\mathcal{O}(K)$	Group of regular $K \times K$ orthogonal (rotation or roto-inversion) matrices and its associated $K \times (K - 1)$ -dimensional subspace	23
$\mathcal{SO}(K)$	Group of regular $K \times K$ special orthogonal (rotation) matrices and its associated $K \times (K - 1)$ -dimensional subspace	23

Other symbols

δ_{pq}	Kronecker symbol: $\delta_{pq} = 1$ if $p = q$, 0 otherwise	153
$\text{cum}(.)$	Cumulant	12
\mathcal{A}_X	Support set of X	6
$\Omega(X)$	Support set of X	6
$\Omega[p_X]$	Support set of X where X is any rv admitting p_X for pdf	6
$\bar{\Omega}(X)$	Convex hull of $\Omega(X)$	42
$R(X)$	Range of X	60
$H(.), H[.]$	Shannon's entropy (discrete)	32
$h(.), h[.]$	Shannon's differential entropy (continuous)	33
$\hat{h}[p]$	Shannon's differential entropy (continuous) of a density p estimated using Parzen window estimator and Riemann summation instead of exact integration	100
$\bar{h}[p]$	Shannon's differential entropy (continuous) of the exact density p (or obtained via numerical convolution of exact densities) using Riemann summation instead of exact integration	110
$\mathcal{H}(.), \mathcal{H}[.]$	Upperbound of Shannon's differential entropy	109
$D(. .)$	Divergence measure between pdf	6
$\text{KL}[. .]$	Kullback-Leibler divergence	15
$\text{KL}[.]$	Mutual information (continuous)	16
$JS_{\pi}(p_1, \dots, p_N)$	Generalized Jensen-Shannon between densities p_n with prior π_n	
	172	
$H_r(.), H_{r,\Omega}(.)$	Discrete Rényi's entropy	40
$h_r(.), h_{r,\Omega}(.)$	Rényi's entropy	40

$h_{r,\bar{\Omega}}(\cdot)$	Extended Rényi's entropy	41
$\bar{h}_r[p]$	Rényi's entropy of the exact density p (or obtained via numerical convolution of exact densities) using Riemann summation instead of exact integration	74
$\sharp[\cdot]$	Cardinal operator	7
$\widetilde{\operatorname{argmax}}(f)$	Set of points in $\operatorname{dom}(f)$ locally maximizing function f	58
$\widetilde{\operatorname{argmin}}(f)$	Set of points in $\operatorname{dom}(f)$ locally minimizing function f	58
$\mathcal{C}(\mathbf{b})$	D-BSS Contrast function	20
$\mathcal{C}(\mathbf{B})$	S-BSS Contrast function if $\mathbf{B} \in \mathcal{M}(K)$, P-BSS contrast function if $\mathbf{B} \in \mathcal{M}^{P \times K}$	20
$\mathcal{C}_h(\cdot)$	Shannon's entropy-based deflation, simultaneous or partial BSS contrast function	51
$\mathcal{C}_R(\cdot)$	Range-based deflation, simultaneous or partial BSS contrast function	61

INTRODUCTION

The XX-th century has been marked by the birth of a new area in mathematical science: *communication*. Its main contributor is certainly, up to now, its inventor: Claude E. Shannon. Since 1948 Shannon's seminal paper entitled "A mathematical theory of communication", statistical signal processing has revealed to be an unmissable tool in the electrical engineering community. In his paper, Shannon has suggested to model the communication problems by stochastic processes. He was the first to ask (and answer) the following question: "Can we define a quantity which will measure, in some sense, how much information is contained in a message that has been emitted by a so-called information source?". He showed that the entropy function, a concept due to Boltzmann and already widely used in physics, fulfills the above requirement.

Nowadays, this quantity plays a fundamental role in, among others, physical chemistry, physics, mechanics, cosmology and obviously, signal processing. A considerable and unexpected power of the entropy is its intuitive interpretation: it tells us about the randomness of a process, it is the key point of the second law of thermodynamics and, even more surprisingly, it is the more common way to define the direction of time! [see Greene, 2005, Reinchenbach, 1984, Gell-Mann and Hartle, 1996]

Shannon has modeled communication systems as information sources (physical entities) that are sending a given message (e.g. an audible sound with a specific meaning). The latter is converted in a signal (most often an acoustic or electrical signal) via a transmitter (e.g. a microphone), and sent to a receiver via a propagation medium (electric wire, air, ...). The received message might be contaminated by some additional noise. The signal is then converted by the receiver (a.o. loudspeakers) to a message (again an audible sound), which should be ideally very close to the original message sent by the source. Finally, the message is forwarded to a destination, a second physical entity to which the original message had to be sent.

Nowadays, the world is full of emitters and receivers, between which signals are transmitted and, unfortunately, interferences are observed. This tendency to spread sensors around us should continue to increase, as recently pointed out by Martin Vetterli in the plenary talk *Distributed signal processing for sensor networks* given at EUSIPCO 2006 (such a popular kind of sensors are the Berkeley

motes). Therefore, an extension of the “one-to-one” Shannon’s communication model, including several physical entities that are communicating simultaneously in a common medium at same time, is needed. This specific area of signal processing considers the additional problem that each receiver records a mixture of the original messages; the channel “perturbations” are related to each other. Assuming a specific propagation model for the messages, we would like to know if it is possible to recover the original information that has been sent by the individual emitters knowing the recordings that have been collected by the sensors. Clearly this is possible when “source coding” can be managed (think about radio-emission, mobile phones, etc). For example, you can select a specific radiophonic program if the emission bandwidths of the radio stations are different by using frequency or amplitude modulation/demodulation and filtering. But this is not always possible.

For instance, the problem that consists in separating acoustic speech signals or astrophysical signals does not enter this framework as we have no way to adapt the source coding; we have no way to access (and thus code) the source signals. They have to be processed directly, as they are sent in the medium by the physical sources. This problem, called “source separation”, is a critical issue. When no (actually, very weak) information is known about the sources and/or the propagation medium, the problem is refer to as *Blind Source Separation problem* (BSS); the separation is based on the signals collected at the receivers *only*. This often occurs in biomedical applications.

BSS seems to be untractable: how can we recover source signals without knowing them neither the mixing system? This is possible for the human being: even if some people are simultaneously speaking in the same room, we can understand what a specific person is saying without knowing in advance anything about the speaker or his message, or knowing the mathematical description of the message propagation through the ambient medium. But we can exploit additional information which is not available to a machine. For example, assume that the four source signals, shown in the left of Fig. I.1., are waveforms of sound signals linearly mixed to produce the mixture signals plotted in the middle of the figure. We assume that neither the mixing coefficients nor the source signals are known; only the mixtures are available.

If one listens to the four mixed sound signals one shall probably recognize at the end that each of them is a mixture of the first notes of the James Bond “Goldeneye” theme, the “Ketchup song”, the sentence pronounced by a female “si vous comprenez cette phrase, votre algorithme de separation fonctionne correctement” (said twice) and a noisy sound. Transcribing each of these signals is another thing but, still, one can make some use of the information emitted by these sources as they can be “separated”. For the computer, however, none of these signals has a specific meaning: they are nothing but electrical signals. Consequently, will a *sufficiently intelligent* machine be able to separate them, too? The answer is yes: according to 1994 Comon’s paper “Independent Component Analysis, A New Concept?”, the source messages can be recovered under mild assumptions by globally maximizing a so-called *contrast function*. More con-

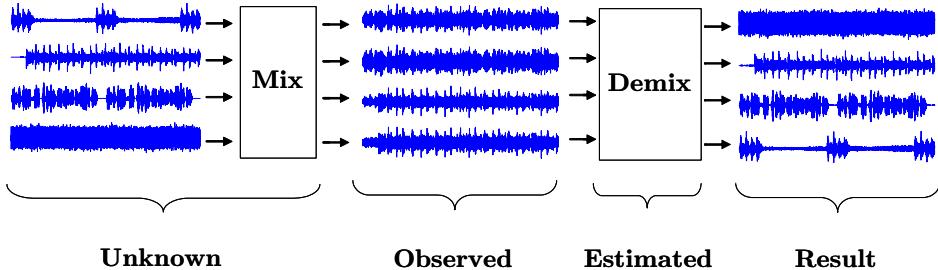


Figure I.1. Blind processing of mixed signals (middle), which are linear mixtures of source signals (left) produces estimated sources (right) by estimating the demixing system.

cretely, maximizing a suitable contrast function based on the signals shown in the middle of Figure I.1. yields the demixing coefficients: the estimated sources, whose temporal waveform are shown in the right of Figure I.1. resemble the waveforms of the original “soundtracks”, possibly up to a permutation.

The above considerations briefly specify the context of this thesis, which deals with these two exciting areas of electrical engineering: the use of entropy (and its generalized form due to Alfred Rényi) in the framework of blind source separation. More specifically, we shall analyze the contrast properties of entropy-related criteria: are they all contrast functions? How “suitable” are there from the optimization point of view?

In order to fix the framework, Chapter 1 presents BSS and its relationship to Independent Component Analysis. Entropy, which is nothing but a measure of information, is shown to be a promising criterion for BSS. Based on the defined concepts, the issue of the thesis can then be more clearly stated.

In Chapter 2, the critical points of entropies are studied in the BSS context, and some tools for building contrast functions from this concept are given. In addition to the state-of-the-art, some additional results due to Dinh-Tuan Pham are provided regarding Shannon’s entropy based on a Taylor’s development. Next, the *support* (or *range*) function is proved to be a contrast based on the Brunn-Minkowski inequality, a similar form of Entropy Power Inequality, a well-known theorem regarding Shannon’s entropy. These two approaches are seen to be particular cases of Rényi’s entropy. It is then natural to try to develop a more general theory that would unify all the approaches based on Rényi’s entropy. Unfortunately, we conclude that generally speaking, some conditions that are difficult to check have to be met when general Rényi’s entropy is used in the BSS context (without restriction on the value of Rényi’s exponent).

However, a major problem remains in the related literature, even if a very simple model is assumed for the propagation medium: “how can we find the global maximizer of a function?” When algebraic methods corresponding to the criteria are available, the problem can be easily managed, but unfortunately,

this is not the case in the BSS applications when “exact” entropy-based contrast functions are used. Then, some optimization algorithms (like gradient-ascent techniques) have been proposed for adaptively locally maximizing the contrast, *without guarantee that the local maximum found will be the global (seeked) one.* The problem would be much easier to solve (and actually vanishes) if the local maxima of the contrast function are all global ones (each corresponding to the separation of the original sources), or if all the local maxima of the contrast correspond to a satisfactory solution of the BSS problem ! This so-called *discriminacy* property would be very useful; it would give confidence in the results returned by the optimization algorithm. This ideal possible behavior was the guideline and motivation of Nathalie Delfosse and Philippe Loubaton in 1995 detailed in their key paper “Adaptive Blind Separation of Sources: A Deflation Approach”. In this work, a kurtosis-based contrast function is proved to benefit from this nice advantage, when the sources are extracted iteratively, one by one.

One of the main topics of this thesis is to tackle the lack of knowledge on the BSS entropic contrasts; this is done in Chapter 3. It is proved that in specific situations, the entropic contrast might suffer from the existence of spurious local maxima, in which the optimization algorithm may be stuck. A same conclusion is drawn regarding the minimum support approach. Then, a slight variant is proposed, the minimum range approach, which benefits from the discriminacy property *whatever the extraction scheme*, that is even if the signals are not estimated sequentially like in the Delfosse & Loubaton method. This result deserves to be emphasized because of its uniqueness; to our knowledge, this contrast is the only one used in simultaneous separation for which the discriminacy property has been established, up to now.

Therefore, Chapter 4 logically focuses on that criterion. A geometric interpretation is given and the practical problems related to range estimation are discussed. This method also proves efficient to separate correlated signals (such as images with common shape: human face pictures, landscapes, etc.) by slightly modifying the optimization algorithm in order to relax the rigid orthogonalization constraint. Finally, the minimum range method, which applies only to double-bounded signals (with finite range) is extended to signals bounded on one side only. This method proves to perform well on the “IEEE MLSP 2006 data analysis competition” data set; this is detailed in the last section of the thesis.

CHAPTER 1

BLIND SOURCE SEPARATION AND ITS RELATIONSHIP TO INDEPENDENT COMPONENT ANALYSIS AND INFORMATION MEASURES

Abstract. This chapter aims at giving the context of the work as well as defining mathematical and conceptual notions needed in the sequel. It is also explained why the concept of independence measure is not *the panacea* to solve the BSS problem even when the sources are independent, and that it is of interest to sketch this problem in the context of information theory. From this viewpoint, entropy-based approaches are unified using Rényi entropies: they are all information measures. The issues of this thesis are then clarified.

Contribution. We define the concepts of “non-mixing matrices” and “(sub)PD-equivalency”, that have relationship and are closely linked to the BSS problem. The *contrast function* definition due to Comon in 1994 is extended to define *simultaneous*, *deflation* and *partial* contrast functions; they can be plugged into simultaneous, deflation and partial separation schemes, respectively. We propose an information-theoretic approach to ICA, based on the concept of “information measures” due to Hartley. This extends the usual minimum output-dependence approach, which is shown to be meaningful for BSS in a simultaneous separation scheme only. This viewpoint suggests that the general class of Rényi entropies would deserve to be further analyzed for BSS (they are widely used, but the underlying motivation remains subjective and misunderstood). We explain why the central limit theorem is a good intuitive approach, though not a formal proof for justifying the use of Shannon’s entropy. Rather, information-theoretic inequalities such as the entropy power inequality (first conjectured by Shannon

in 1948, and latter formally proved by Stam in 1959) is used to that aim. An extended form of Rényi's entropy is also proposed to further include the range approach in the class of informative criteria.

Part of the work presented in this chapter was published in JP1 (see Appendix B).

Organization of the chapter. In the first section of this chapter, the Blind Source Separation (BSS) task will first be sketched as well as the mathematical model and corresponding assumptions. Section 1.2 introduces the well-known method of Independent Component Analysis (ICA); based on Comon's identifiability Theorem, connections between BSS and ICA are emphasized. Section 1.3 gives a non-exhaustive list of examples of independence measures that are/could be used in ICA. In spite of the relationships between BSS and ICA, they cannot be seen to be rigorously equivalent problems even under the assumption of independent sources, especially when deflation or partial separation schemes are considered. Therefore, Section 1.4 presents the three possible extraction schemes as well as a corresponding general formulation in terms of the optimization of a so-called contrast function. The dimension of the definition domain of these functions (i.e. of the search space) can be reduced thanks to the so-called whitening preprocessing, where the parameter space is the group of (semi)-orthogonal matrices or the set of orthonormal vectors, and the corresponding contrasts are named orthogonal contrasts; this is presented in Section 1.5. Section 1.6 reminds gradient-ascent rules designed for the extractions schemes of the contrasts, as often, algebraic techniques are not available for the maximization of the contrast functions. Even if the source independence assumption is still needed, Section 1.7 proposes a different way to consider the BSS problem than the usual ICA. Rather than the output dependence, the output "complexity" is minimized, where the complexity measure can be seen interestingly to be linked to information measure. Finally, Section 1.8 introduces the objectives of this thesis, which is mainly the analysis of the contrast properties of the class of Rényi's information measures.

1.1 SOURCE SEPARATION: MOTIVATION AND MODELS

One of the main reasons for justifying the impressive development of BSS-related techniques is the wide range of applications. Let us consider the BSS problem as first sketched in the introduction: the aim is to recover source signals from mixtures of them. This problem is often illustrated by the so-called cocktail party problem. Assume that K persons are simultaneously speaking in a room, as it often occurs in a cocktail party, where in several smaller groups of persons, one person is speaking. Assume further that a *sufficient number of microphones*, say N , are located at different places in the room. Obviously, each microphone does not record an individual speech, but rather a kind of *superimposition* of the *sound*

messages; the recording depends on the location of the microphones. The BSS task is to find a suitable method that would process the microphone recordings and give, as output, the original speeches. The cocktail party problem is often used as an illustrative and comprehensive application of the BSS. However, many other real-world applications can be modelled as a BSS problem. Here is a non-exhaustive list (some references can be found in [Hyvärinen et al., 2001, Cichoki and Amari, 2002, Pham and Jutten, 2003, Antoniadis et al., 2001])

- Audio processing
- Power plants monitoring
- Seismic and astrophysical signals analysis
- Denoising
- Biomedical signal analysis
- Image processing
- ...

All the above applications share the same model: K sources $\mathbf{S}_1(t), \dots, \mathbf{S}_K(t)$, emitted by a physical entity, have to be recovered via N mixtures of them, say $\mathbf{X}_1(t), \dots, \mathbf{X}_N(t)$; these signals are random variables (t is implicitly assumed to be discrete). The resulting mixtures obviously depend on the original sources, their respective location compared to the microphones, the propagation medium, and the characteristics of the sensors. Such a model can be written in a quite simple way.

Denote, at time t , the sources vector by $\mathbf{S}(t) = [\mathbf{S}_1(t), \dots, \mathbf{S}_K(t)]^T$, the recordings by $\mathbf{X}(t) = [\mathbf{X}_1(t), \dots, \mathbf{X}_N(t)]^T$. They are linked by the following relation:

$$\mathbf{X}(t) = \mathcal{F}((\mathbf{S}(t), \mathbf{S}(t-1), \dots), t) , \quad (1.1)$$

where $\mathcal{F}(\cdot, t)$ denotes the mixing system at time t . With these notations, we can sketch a first definition of the general BSS task.

Definition 1 (General BSS) *Assuming that a vector $\mathbf{X}(t)$ of N mixtures is known and that $\mathbf{X}(t) = \mathcal{F}(\mathbf{S}(t), t)$, where $\mathcal{F}(\cdot, t)$ is the unknown mixing system at time t and $\mathbf{S}(t)$ is a vector of K source signals, the BSS task is to blindly find a demixing system $\mathcal{G}(\cdot, t)$ such that*

$$\mathbf{S}(t) = \mathcal{G}(\mathbf{X}(t), t) . \quad (1.2)$$

The usual approach of classical signal processing would be to model the mixing system $\mathcal{F}(\cdot, t)$ by using the physical specificities of the propagation medium

and of the sensors and then, to invert this system. However, this is a tedious task, since it requires to have a lot of information on both the mixture scheme and the sensor specificities, which is often lacking (just think about biomedical applications).

Rigourously speaking, BSS should refer to the problem of separating the sources whatever they are and whatever the mixture scheme. Unfortunately, this is not possible in practice: some assumptions have to be made. Consequently, a new challenge consists in making some assumptions on the mixing system, and then trying to estimate the parameters of this model according to the estimation theory, i.e. by using e.g. maximum likelihood techniques. We fix a specific model for the demixing system $\mathcal{G}(\cdot|\Theta, t)$ where Θ denotes the (set of) parameter(s) involved in the model, belonging to some parameter space \mathcal{T} , i.e. we assume that \mathcal{G} has a specific form, and that there exists a (set of) parameter(s) $\Theta^* \in \mathcal{T}$ such that $\mathcal{G}(\cdot|\Theta^*, t) = \mathcal{F}^{-1}(\cdot, t)$.

In the BSS community, the acronym *blind* has thus a specific meaning: implicitly, mild assumptions are made. Several assumptions on the pair [mixture scheme, sensor] can be drawn. One can deal with linear, convolutive, post-nonlinear, or convolutive post-nonlinear mixture schemes (see e.g. the monograph [Hyvärinen, Karhunen, and Oja, 2001]). This thesis focuses on the simplest (but also most used) BSS model. It is assumed that the mixing system is linear, time invariant and instantaneous: in other words,

$$\mathcal{F}(S(t), t) = AS(t) , \quad (1.3)$$

where A is an $N \times K$ mixing matrix.

Consequently, the demixing system \mathcal{G} also reduces to a single matrix B , and $\mathcal{G} \circ \mathcal{F}$ is the identity mapping if and only if $BA = I_K$ with I_K the identity matrix of order K .

Then, $\Theta^* = B^*$ with $B^*A = I_K$ and $\mathcal{T} = \mathbb{R}^{K \times N}$. Since the mixing system is memoryless and time-invariant, we can drop the time index t . The new aim of BSS is thus the following:

Definition 2 (Linear, instantaneous, time-invariant BSS) *Let us assume a mixture model $\mathcal{F}(S) = AS$ for the mixtures, where both the mixing matrix and the vector of sources are unknown. The goal of BSS is to find a demixing system $\mathcal{G}(X) = BX$ such that $\mathcal{G} \circ \mathcal{F}(S) = S$, that is to find the demixing matrix B such that $BA = I_K$.*

The above definition states the problem, but does not ensure that the sources are *separable*, yet. In other words, we have no guarantee that without further hypotheses, we are indeed able to find such a demixing matrix B^* from the mixture vector X only. Actually, the above assumptions on the mixing system are not sufficient to ensure separability; we need further hypotheses on the sources. As for the mixing system, a wide variety of hypotheses can be made, each one yielding to a specific method solving the BSS problem: the sources can e.g. be independent and identically distributed (i.i.d.) or, on the contrary, with temporal structure; they can be bounded, take only positive values, etc.

When the assumptions are much stronger, as it is the case when additional information is available on the mixing system (the entries of the mixing matrix are constrained to be positive coefficients, some of them are known, ...) or on the sources (some of the sources pdf are available, sparse, ...), the problem is often referred to as *semi-blind* source separation. For instance, these approaches include among other Bayesian methods, support-based criteria, application-driven information such as in audio source separation (sparsity, time-frequency masking, etc).

In spite of the terminology, the general BSS problem is thus untractable. So why reserving this term to a problem which is impossible to solve in practice? Rather, the BSS problem usually refers to Definition 2 (p. 4), with additional assumption on the sources and still further constraints on the mixing system. They are summarized in the following list.

Assumptions on the mixing matrix \mathbf{A}

\mathcal{A}_1 \mathbf{A} is a square matrix of size $K \times K$: $\mathbf{A} \in \mathbb{R}^{K \times K}$.

\mathcal{A}_2 \mathbf{A} is invertible, and thus of full rank : $\text{rank}(\mathbf{A}) = K$.

- The joint assumption $\mathcal{A}_{12} = \mathcal{A}_1 \wedge \mathcal{A}_2$ is equivalent to the requirement $\mathbf{A} \in \mathcal{M}(K)$, where $\mathcal{M}(K)$ is the General Linear group of degree K , defined by :

$$\mathcal{M}(K) \doteq \{\mathbf{M} \in \mathbb{R}^{K \times K} : \text{rank}(\mathbf{M}) = K\} . \quad (1.4)$$

It should be stressed that \mathcal{A}_1 is actually too restrictive. Clearly, if \mathbf{A} is not square, it is not invertible, but if $N > K$ and $\text{row-rank}(\mathbf{A}) = K$ it is still possible to recover the K sources. Indeed, if $N > K$ the mixture scheme is said to be *overdetermined* (or *undercomplete*), and it suffices to “discard” some components of the mixture vector \mathbf{X} that can be generated by a linear combination of other rows of the mixing matrix. Clearly, $N - K$ such mixtures exist as the rows of \mathbf{A} span a K -dimensional space.

Finding these redundant mixtures is a priori not an easy task. It is possible, however, to use ad-hoc preprocessing methods, such as Principal Component Analysis (PCA), to project the mixture data from a N -dimensional space to a K dimensional one without any loss of information [Hyvärinen, Karhunen, and Oja, 2001] (only the redundancies vanish). This procedure yields a new K -dimensional vector \mathbf{X} (which is not simply composed of K of the N components of the original mixture vector, but rather of linear combinations of them), which could have been generated by a full-rank $K \times K$ mixing matrix, which is clearly invertible. Therefore, such a preprocessing ensures that \mathcal{A}_1 now holds.

More precisely, let us define the statistical expectation of a function f of a random variable (r.v.) \mathbf{X} with respect to its density $p_{\mathbf{X}}$ by

$$\text{Ex}[f(\mathbf{X})] \doteq \sum_{x \in \mathcal{A}_{\mathbf{X}}} p_{\mathbf{X}}(x)f(x) \text{ if } \mathbf{X} \text{ is a discrete r.v.} \quad (1.5)$$

$$\text{Ex}[f(\mathbf{X})] \doteq \int_{x \in \Omega(\mathbf{X})} p_{\mathbf{X}}(x)f(x)dx \text{ if } \mathbf{X} \text{ is a continuous r.v.} \quad (1.6)$$

In the last definitions, \mathcal{A}_X is the countable alphabet of X and $\Omega(X)$ is the support set of X , the latter is defined in the one-dimensional space as

$$\Omega(X) \doteq \{x \in \mathbb{R} : p_X(x) > 0\} . \quad (1.7)$$

Note that if we extend the pdf definition domain to \mathbb{R} such that $p_X(x) = 0$ for $x \in \mathbb{R} \setminus \Omega(X)$, then

$$E[f(X)] = \int_{\mathbb{R}} p_X(x) f(x) dx , \quad (1.8)$$

where the subscript X is dropped for short when no confusion is possible.

PCA consists in projecting the data onto the eigenvectors of

$$\text{Cov}[X] \doteq E[XX^T] - E^2[X] , \quad (1.9)$$

(the covariance matrix of the noise-free observed mixtures) associated with the K non-zero eigenvalues among the N obtained via e.g. eigenvalue decomposition (EVD) of $\text{Cov}[X]$. It only cancels the linear redundancies contained in the data by projecting them onto the basis formed by the eigenvectors of $\text{Cov}[X]$ which is orthogonal because the covariance matrix is symmetric (this yields, in addition, uncorrelated signals, see Section 1.5). Then, we shall assume $N = K$ even if the more general case $N \geq K$ can be easily managed.

Assumptions on the sources

\mathcal{A}_3 The sources are identically distributed; for each source index $i \in \{1, \dots, K\}$, the probability density function (pdf) $p_{S_i(t)}(S_i(t))$ of $S_i(t)$ does not depend on the time index t (and this index can thus be omitted when it is not necessary).

\mathcal{A}_4 The sources are zero-mean: $E[S_i] = 0$, $i \in \{1, \dots, K\}$.

\mathcal{A}_5 The sources are mutually independent:

$$p_S(S) = p_{S_1, \dots, S_K}(S_1, \dots, S_K) = \prod_{i=1}^K p_{S_i}(S_i) . \quad (1.10)$$

\mathcal{A}_6 There is at most one Gaussian source; noting a divergence measure between densities as $D(p||q)$ satisfying $D(p||q) \geq 0$ with equality if and only $p = q$ almost everywhere we have

$$\sharp[\{i \in \{1, \dots, K\} : D(p_{S_i}||\phi_{S_i}) = 0\}] \leq 1 , \quad (1.11)$$

where ϕ_{S_i} is the zero-mean Normal pdf with same variance $\sigma_{S_i}^2$ as S_i

$$\phi_{S_i} = \frac{1}{\sqrt{2\pi\sigma_{S_i}^2}} e^{-\frac{x^2}{2\sigma_{S_i}^2}} , \quad (1.12)$$

and $\#[.]$ is the cardinal operator.

From \mathcal{A}_{12} , $\mathbf{AA}^{-1} = \mathbf{I}_K$ where \mathbf{A}^{-1} is unique. We are now able to state the BSS problem definition, which is thus not blind at all actually, as understood by the scientific community:

Definition 3 (BSS) *Assuming that $\mathbf{X} = \mathbf{AS}$ where $\mathbf{A} \in \mathcal{M}(K)$ and the vector of sources are both unknown (up to the assumptions $\mathcal{A}_3 - \mathcal{A}_6$), find a demixing matrix $\mathbf{B} \in \mathcal{M}(K)$ such that $\mathbf{BA} = \mathbf{I}_K$.*

In 1994, Comon has shown that in the BSS framework defined in Def. 3 (p. 7), $\mathbf{B} = \mathbf{A}^{-1}$ is identifiable, but only up to some indeterminacies [Comon, 1994]. Such a matrix can be found through an ICA, as detailed in the next section.

1.2 INDEPENDENT COMPONENT ANALYSIS : AN EFFICIENT TOOL FOR BLIND SOURCE SEPARATION

Blind source separation has been defined in Def. 3 but, still, we have no way to recover the original sources, i.e. we need some tool for estimating $\mathbf{B} \approx \mathbf{A}^{-1}$. Independent Component Analysis (ICA) aims at recovering independent components from a random vector; it is thus apparently a somewhat different problem; however, both BSS and ICA are closely connected to each other. We shall first review some independence-related concepts before illustrating the relationships between ICA and BSS.

1.2.1 PD-equivalency and Non-mixing matrices

Let $\mathcal{P}(K)$, $\mathcal{D}(K)$ be the subgroups¹ of permutation and diagonal invertible matrices, with thus $\mathcal{P}(K) \subset \mathcal{M}(K)$ and $\mathcal{D}(K) \subset \mathcal{M}(K)$. We now define two important other sets of matrices.

Similarly to Eq. (1.4), we define by $\mathcal{M}^{P \times K}$ the set of full row-rank $P \times K$ matrices:

$$\mathcal{M}^{P \times K} \doteq \{\mathbf{M} \in \mathbb{R}^{P \times K} : \text{row-rank}(\mathbf{M}) = P, P \leq K\} . \quad (1.13)$$

Important subsets of $\mathcal{M}(K)$ and $\mathcal{M}^{P \times K}$ are the sets of non-mixing matrices.

Definition 4 (Set of non-mixing matrices) *The set of non-mixing matrices of order K is defined as the set of $K \times K$ matrices having a single non-zero element per row and per column*

$$\mathcal{W}(K) \doteq \{\mathbf{M} : \exists \boldsymbol{\Lambda} \in \mathcal{D}(K), \boldsymbol{\Pi} \in \mathcal{P}(K), \mathbf{M} = \boldsymbol{\Lambda} \boldsymbol{\Pi}\} . \quad (1.14)$$

¹It is easy to observe that these sets have indeed the group structure under matrix multiplication satisfying closure, associativity, inverse and identity.

It is easily seen that $\mathcal{W}(K)$ forms a group, also called the group of monomial matrices.

A $K \times K$ matrix \mathbf{M} is said to be non-mixing if $\mathbf{M} \in \mathcal{W}(K)$. Similarly, a $P \times K$ rectangular matrix, $P \leq K$, is said to be non-mixing if each of its row is a distinct row of a given matrix $\mathbf{M} \in \mathcal{W}(K)$ (it has a single non-zero element per row). If we define $\underline{\Pi}$ as a matrix composed of P distinct rows of $\Pi \in \mathcal{P}(K)$, then the set $\mathcal{W}^{P \times K}$ of $P \times K$ non-mixing matrices is

$$\mathcal{W}^{P \times K} \doteq \{\mathbf{M} : \exists \Lambda \in \mathcal{D}(P), \mathbf{M} = \Lambda \underline{\Pi}\}. \quad (1.15)$$

This set is the set of rectangular matrices with a single non-zero element per row and at most one per column; it satisfies $\mathcal{W}^{K \times K} = \mathcal{W}(K)$.

Example 1 (Non-mixing matrices) From the above definition:

$$\begin{bmatrix} -2/3 & 0 & 0 \\ 0 & 0 & 4 \\ 0 & 0.3 & 0 \end{bmatrix} \in \mathcal{W}(3), \text{ but } \begin{bmatrix} 0 & 1 & 0 \\ 0 & .2 & 1/4 \\ 6 & 0 & 0 \end{bmatrix} \notin \mathcal{W}(3)$$

and

$$\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 3 & 0 & 0 \end{bmatrix} \in \mathcal{W}^{3 \times 4}, \text{ but } \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \notin \mathcal{W}^{2 \times 3}$$

Let us now define the PD-equivelency operator.

Definition 5 (PD-equivelency) A matrix \mathbf{M} is said to be PD-equivalent to \mathbf{B} , noted $\mathbf{M} \sim \mathbf{B}$, if there exists two matrices $\Pi \in \mathcal{P}(K)$, $\Lambda \in \mathcal{D}(K)$ such that $\mathbf{B} = \Lambda \Pi \mathbf{M}$ or, equivalently, if there exists a non-mixing matrix $\mathbf{N} \in \mathcal{W}(K)$ such that $\mathbf{B} = \mathbf{N} \mathbf{M}$.

Observe that the PD-equivelency operator is, by definition, scale and permutation invariant but it also further satisfies three important properties, summarized in Lemma 2. The following lemma is useful for proving Lemma 2.

Lemma 1 For any product $\Pi \Lambda$ where $\Pi \in \mathcal{P}(K)$, $\Lambda \in \mathcal{D}(K)$, there exists $\Lambda' \in \mathcal{D}(K)$ such that $\Pi \Lambda = \Lambda' \Pi$. Conversely, for any product $\Lambda \Pi$, there exists $\Lambda' \in \mathcal{D}(K)$ such that $\Lambda \Pi = \Pi \Lambda'$.

Proof: We only show the first claim; the converse is proved in the same way. By definition of permutation matrices, any i -th row of Π corresponds to a j -th row \mathbf{e}_j of the identity matrix \mathbf{I}_K . Let us denote by $j(i)$ the column index of the single non-zero element of the i -th row of the permutation matrix Π (that is the indice j of the row of \mathbf{I}_K corresponding to the i -th row of Π); for all $1 \leq i \leq P$, $[\Pi]_{ij} \neq 0$ if and only if $j = j(i)$. The matrix $\Pi \Lambda$ is obtained by replacing the i -th row of Λ by the $j(i)$ -th one. On the other hand, $\Lambda \Pi$ is obtained by replacing the i -th column of Λ by the $j'(i)$ -th one, where $j'(i)$ is defined as $j(i)$ but with Π^T instead of Π : for all $1 \leq i \leq P$, $[\Pi^T]_{ij} \neq 0$ if and only if $j = j'(i)$.

The equality $\mathbf{\Pi}\mathbf{\Lambda} = \mathbf{\Lambda}'\mathbf{\Pi}$ is equivalent to $\mathbf{\Pi}\mathbf{\Lambda}\mathbf{\Pi}^{-1} = \mathbf{\Lambda}'$ with $\mathbf{\Pi}^{-1} = \mathbf{\Pi}^T$ because $\mathbf{\Pi}$ is a permutation matrix. Let us now prove that $\mathbf{\Pi}\mathbf{\Lambda}\mathbf{\Pi}^T$ is in the set $\mathcal{D}(K)$ of $K \times K$ diagonal matrices. As explained above, the left-multiplication of $\mathbf{\Lambda}$ by a permutation matrix moves the (k, l) -th element to the $(j(k), l)$ -th place. The right multiplication of a matrix by $\mathbf{\Pi}$ (resp. $\mathbf{\Pi}^T$) moves the (k, l) -th element to the $(k, j'(l))$ (resp. $(k, j(l))$). Hence, the elements of $\mathbf{\Lambda}'$ are related to those of $\mathbf{\Lambda}$ by the relation $[\mathbf{\Lambda}]_{kl} = [\mathbf{\Lambda}']_{j(k), j(l)}$ which shows that $\mathbf{\Lambda}' \in \mathcal{D}(K)$ since $\{j(i), i \in \{1, \dots, K\}\}$ forms a permutation of $\{1, \dots, K\}$ and therefore $j(k) = j(l)$ if and only if $k = l$.

□

Lemma 2 *The following holds true for the PD-equivalency operator.*

- *it is transitive: if $\mathbf{M}_1 \sim \mathbf{M}_2$ and $\mathbf{M}_2 \sim \mathbf{M}_3$, then $\mathbf{M}_1 \sim \mathbf{M}_3$.*
- *it is invariant under right multiplication: if $\mathbf{M}_1 \sim \mathbf{M}_2$, then $\mathbf{M}_1\mathbf{M}_3 \sim \mathbf{M}_2\mathbf{M}_3$.*
- *it is symmetric: For any pair of matrices $\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}(K)$,*

$$\mathbf{M}_1 \sim \mathbf{M}_2 \Leftrightarrow \mathbf{M}_2 \sim \mathbf{M}_1 . \quad (1.16)$$

Proof: By definition of PD-equivalency, there exists $\mathbf{\Lambda}_1, \mathbf{\Lambda}_2 \in \mathcal{D}(K)$, $\mathbf{\Pi}_1, \mathbf{\Pi}_2 \in \mathcal{P}(K)$ such that $\mathbf{M}_1 = \mathbf{\Lambda}_1\mathbf{\Pi}_1\mathbf{M}_2$ and $\mathbf{M}_2 = \mathbf{\Lambda}_2\mathbf{\Pi}_2\mathbf{M}_3$, i.e. $\mathbf{M}_1 = \mathbf{\Lambda}_1\mathbf{\Pi}_1\mathbf{\Lambda}_2\mathbf{\Pi}_2\mathbf{M}_3$. But by Lemma 1, there exists $\mathbf{\Lambda}_3$ s.t. $\mathbf{\Lambda}_2\mathbf{\Pi}_2 = \mathbf{\Pi}_2\mathbf{\Lambda}_3$. Thus, noting that $\mathbf{\Pi}_3 = \mathbf{\Pi}_1\mathbf{\Pi}_2 \in \mathcal{P}(K)$ and $\mathbf{\Lambda}_3\mathbf{\Lambda}_3$ can be rewritten as $\mathbf{\Lambda}_4\mathbf{\Pi}_3$, $\mathbf{M}_1 = \mathbf{\Lambda}_1\mathbf{\Lambda}_4\mathbf{\Pi}_3\mathbf{M}_3$ yielding $\mathbf{M}_1 \sim \mathbf{M}_3$ since $\mathbf{\Lambda}_1\mathbf{\Lambda}_4 \in \mathcal{D}(K)$. This proves the first property.

The second property is trivial because $\mathbf{M}_1\mathbf{M}_3 \sim \underbrace{\mathbf{\Lambda}\mathbf{\Pi}\mathbf{M}_1}_{\mathbf{M}_2}\mathbf{M}_3$ and of Lemma 1.

Finally, the last property results from the group structure of $\mathcal{P}(K)$ and $\mathcal{D}(K)$ (each element of the sets of permutation and diagonal matrices is invertible and the inverse is in the respective set: if $\mathbf{M}_1 = \mathbf{\Lambda}_1\mathbf{\Pi}_1\mathbf{M}_2$, we have $\mathbf{M}_2 = \mathbf{\Pi}_2\mathbf{\Lambda}_2\mathbf{M}_1$, where $\mathbf{\Pi}_2 = \mathbf{\Pi}_1^{-1}$ and $\mathbf{\Lambda}_2 = \mathbf{\Lambda}_1^{-1}$). The property is shown by using Lemma 1.

□

Definition 6 (Invariance under PD-equivalency preserving transforms)
A function $f(\cdot)$ is said invariant under PD-equivalency preserving transforms if $f(\mathbf{B}) = f(\mathbf{M})$ for $\mathbf{M} \sim \mathbf{B}$.

The PD-equivalence operator can be extended as follows.

Definition 7 (SubPD-equivalence) *A matrix $\mathbf{M}_1 \in \mathcal{M}^{P \times K}$ is subPD-equivalent to a square $K \times K$ matrix $\mathbf{M}_2 \in \mathcal{M}(K)$, $P \leq K$, if for all*

$i \in \{1, \dots, P\} \exists j(i) \in \{1, \dots, K\}$ and $\lambda > 0$ such that $[\mathbf{I}_P]_i \mathbf{M}_1 = \lambda [\mathbf{I}_K]_{j(i)} \mathbf{M}_2$ where $[\mathbf{I}_Q]_k$ denotes the k -th row of \mathbf{I}_Q . They are noted $\mathbf{M}_1 \sim_u \mathbf{M}_2$.

In other words, each of the $P \leq K$ rows of \mathbf{M}_1 is proportional to a distinct row of \mathbf{M}_2 (they must be distinct since otherwise $\text{rank}(\mathbf{M}_1) < P$ and $\mathbf{M}_1 \notin \mathcal{M}^{P \times K}$).

Lemma 3 (SubPD-equivalence operator and right multiplication)

Let $\mathbf{M} \in \mathcal{M}^{P \times K}$, $\{\mathbf{M}_1, \mathbf{M}_2\} \subset \mathcal{M}(K)$. Then, if $\mathbf{M} \sim_u \mathbf{M}_1$, $\mathbf{M}\mathbf{M}_2 \sim_u \mathbf{M}_1\mathbf{M}_2$. As a corollary, if $\mathbf{M}\mathbf{M}_1 \sim_u \mathbf{I}_K$, then $\mathbf{M} \sim_u \mathbf{M}_1^{-1}$.

Proof: If $\mathbf{M} \sim_u \mathbf{M}_1$, there exists $\underline{\boldsymbol{\Pi}} \sim_u \boldsymbol{\Pi}$, $\underline{\boldsymbol{\Lambda}} \sim_u \boldsymbol{\Lambda}$ where $\underline{\boldsymbol{\Pi}} \in \mathcal{M}(P)$, $\boldsymbol{\Pi} \in \mathcal{P}(K)$, $\underline{\boldsymbol{\Lambda}} \in \mathcal{D}(P)$ and $\boldsymbol{\Lambda} \in \mathcal{D}(K)$ such that $\underline{\boldsymbol{\Lambda}} \underline{\boldsymbol{\Pi}} \mathbf{M} \sim_u \mathbf{M}_1$. It is always possible to define a matrix $\bar{\mathbf{M}} \in \mathcal{M}(K)$ such that $\boldsymbol{\Lambda} \boldsymbol{\Pi} \bar{\mathbf{M}} \sim \mathbf{M}_1$. But because of Lemma 2, $\boldsymbol{\Lambda} \boldsymbol{\Pi} \bar{\mathbf{M}} \mathbf{M}_2 \sim \mathbf{M}_1 \mathbf{M}_2$ and thus $\underline{\boldsymbol{\Lambda}} \underline{\boldsymbol{\Pi}} \mathbf{M} \mathbf{M}_2 \sim_u \mathbf{M}_1 \mathbf{M}_2$, i.e. $\mathbf{M}\mathbf{M}_2 \sim_u \mathbf{M}_1 \mathbf{M}_2$.

□

Obviously, two PD-equivalent matrices are trivially sub-PD-equivalent and these definitions are equivalent if $P = K$.

We get the following corollary:

Corollary 1 (PD-equivalence and set of non-mixing matrices) *If a matrix \mathbf{M} is (sub)PD-equivalent to a non-mixing matrix, then \mathbf{M} is non-mixing and conversely, if \mathbf{M} is non-mixing, it is (sub)PD-equivalent to a non-mixing matrix.*

The following equivalences hold between PD-equivalence relation and membership to set of non-mixing matrices:

- For a pair $\mathbf{M}_1, \mathbf{M}_2$ of matrices in $\mathcal{M}(K)$: $\mathbf{M}_1 \sim \mathbf{M}_2 \iff \mathbf{M}_1 \mathbf{M}_2^{-1} \in \mathcal{W}(K)$;
- For a pair $\mathbf{M}_1 \in \mathcal{M}^{P \times K}, \mathbf{M}_2 \in \mathcal{M}(K)$: $\mathbf{M}_1 \sim_u \mathbf{M}_2 \iff \mathbf{M}_1 \mathbf{M}_2^{-1} \in \mathcal{W}^{P \times K}$.

Proof: The proof of this corollary is trivial; we deal here with the square case, but the extension to sub-PD-equivelency is straightforward. Let $\mathbf{M}_2 \in \mathcal{W}(K)$ and $\mathbf{M}_1 \sim \mathbf{M}_2$. From Lemma 2, $\mathbf{M}_1 \mathbf{M}_2^{-1} \sim \mathbf{I}_K$ implying that there exists $\boldsymbol{\Lambda} \in \mathcal{D}(K)$, $\boldsymbol{\Pi} \in \mathcal{P}(K)$ such that $\mathbf{M}_1 \mathbf{M}_2^{-1} = \boldsymbol{\Lambda} \boldsymbol{\Pi}$, that is $\mathbf{M}_1 \mathbf{M}_2^{-1} \in \mathcal{W}(K)$. Conversely, if $\mathbf{M} = \mathbf{M}_1 \mathbf{M}_2^{-1} \in \mathcal{W}(K)$, it exists $\boldsymbol{\Lambda} \in \mathcal{D}(K)$, $\boldsymbol{\Pi} \in \mathcal{P}(K)$ such that $\mathbf{M} = \boldsymbol{\Lambda} \boldsymbol{\Pi} \mathbf{I}_K$ which proves $\mathbf{I}_K \sim \mathbf{M}$.

□

As a consequence, the set of non-mixing matrices can alternatively be defined as

$$\mathcal{W}(K) = \{\mathbf{M} \in \mathcal{M}(K) : \mathbf{M} \sim \mathbf{I}_K\} \quad (1.17)$$

and

$$\mathcal{W}^{P \times K} = \{\mathbf{M} \in \mathcal{M}^{P \times K} : \mathbf{M} \sim_u \mathbf{I}_K\} \quad (1.18)$$

1.2.2 Independence and ICA

Let us consider the random variables X and Y that are continuous and for which we can define their probability density functions (pdf) $p_X(x)$, $p_Y(y)$, expectations $E[X]$ and $E[Y]$, variances $\text{Var}[X]$ and $\text{Var}[Y]$, covariance $\text{Cov}[X, Y]$ and Pearson correlation coefficient (correlation, for short) $\text{Corr}[X, Y]$:

$$\text{Var}[X] = E[(X - E[X])^2] \quad (1.19)$$

$$\text{Cov}[X, Y] = E[XY] - E[X]E[Y] \quad (1.20)$$

$$\text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}, \quad (1.21)$$

where the expectation $E[.]$ is defined in Eq. (1.8).

Observe that the statistical definition of independence matches the intuitive one. Indeed, for two independent variables X and Y , we have:

$$p_{X,Y}(x, y) = p_X(x)p_Y(y). \quad (1.22)$$

Noting $p_{X|Y}(x|y)$ the conditional pdf of X given $Y = y$ and using Bayes' rule, it comes that

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}, \quad (1.23)$$

where $p_{X,Y}(x, y)$ is the joint density of (X, Y) at (x, y) . Hence, we obtain that

$$p_{X|Y}(x|y) = p_X(x). \quad (1.24)$$

In other words, there is no information brought on X by knowing that $Y = y$.

Definition 8 (ICA-1) Assume that $\wedge_{i=1}^6 \mathcal{A}_i$ hold. Knowing a K -dimensional vector of observations \mathbf{X} , ICA aims at finding a linear transformation $\mathbf{B} \in \mathcal{M}(K)$ such that the components of $\mathbf{Y} = \mathbf{BX}$ are as independent as possible.

Observe that by contrast to the BSS problem definition as stated in Def. 2 (p. 4), ICA-1 is a tractable problem since it does not involve neither \mathbf{A} nor \mathbf{S} .²

ICA can be seen as an extension of decorrelation. Instead of searching a basis in which the components are decorrelated, we try to find a basis in which the components are made independent. It aims at recovering underlying independent components from the mixture; it is a kind of *higher order, non-linear decorrelation*. Indeed, while decorrelation between two centered variables X and Y is achieved if and only if $E[XY] = E[X]E[Y]$, independence between these variables requires that $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ for all continuous functions

²Note that this must be tempered as in practice: the output densities depend on the pair (\mathbf{A}, \mathbf{S}) . It is thus implicitly assumed that those pdfs can be obtained from the data.

f, g that are zero outside a finite interval [Feller, 1966]. Indeed:

$$\mathbb{E}[f(X)g(Y)] = \int \int f(x)g(y)p_{X,Y}(x,y)dxdy \quad (1.25)$$

$$= \int f(x)p_X(x)dx \int g(y)p_Y(y)dy \quad (1.26)$$

$$= \mathbb{E}[f(X)]\mathbb{E}[g(Y)]. \quad (1.27)$$

In a similar way, while decorrelation between X and Y cancels the second-order cross-cumulants, independence means that all higher-order cross-cumulants are zero, too. In practice, a good approximation of independence consists in having the fourth-order cross-cumulant equal to zero. As a reminder, the fourth-order cumulant of a random vector \mathbf{Y} is defined as

$$\text{cum}_{ijkl}(\mathbf{Y}) \doteq \mathbb{E}[Y_i Y_j Y_k Y_l] - \mathbb{E}[Y_i Y_i]\mathbb{E}[Y_k Y_l] - \mathbb{E}[Y_i Y_k]\mathbb{E}[Y_j Y_l] - \mathbb{E}[Y_i Y_l]\mathbb{E}[Y_j Y_k]. \quad (1.28)$$

1.2.3 ICA and BSS

A similarity between ICA and BSS now arises: both aims at finding a specific demixing matrix; in ICA, “demixing” means recovering independent outputs, while in BSS, we are interested in recovering original sources (supposed to be independent as suggested by \mathcal{A}_5) that have been mixed through the mixing matrix \mathbf{A} . In the following, it will then be assumed that the BSS model defined in Def. 3 (p. 7) holds.

In 1994, Comon has shown that under the above assumptions, the BSS problem can be solved by using ICA. In this section, we shall restrict ourselves to showing that recovering underlying independent components form the mixture leads to identifying the mixing matrix up to some indeterminacies.

The first step to link BSS to ICA is the so-called Darmois-Skitovitch theorem [Darmois, 1953].

Theorem 1 (Darmois-Skitovitch) *Let us suppose that $X_1 = \sum_{i=1}^K \alpha_i S_i$ and $X_2 = \sum_{i=1}^K \beta_i S_i$ where S_1, \dots, S_K are independent rv and $\alpha_j \in \mathbb{R}, \beta_j \in \mathbb{R}, j \in \{1, \dots, K\}$. Then, if X_1 and X_2 are independent, all the S_j such that $\alpha_j \beta_j \neq 0$ are Gaussian (i.e. have a Gaussian pdf).*

This theorem admits a converse (see e.g. [Theis, 2002]).

Theorem 2 (Converse DS) *Let us suppose that $X_1 = \sum_{i=1}^K \alpha_i S_i$ and $X_2 = \sum_{i=1}^K \beta_i S_i$ where S_1, \dots, S_K are independent rv and $\alpha_j \in \mathbb{R}, \beta_j \in \mathbb{R}, j \in \{1, \dots, K\}$. Then, if $\alpha_i \beta_i = 0$ for all $1 \leq i \leq K$, then X_1 is independent from X_2 .*

Based on Theorem 1, Comon has derived the following key theorem.

Theorem 3 (Comon) Let \mathbf{S} be a K -dimensional vector of independent components where at most one of the S_j is Gaussian and $\mathbf{A} \in \mathcal{M}(K)$. Then, setting $\mathbf{X} = \mathbf{AS}$, the following statements are rigorously equivalent:

- X_1, \dots, X_K are pairwise independent;
- X_1, \dots, X_K are mutually independent
- $\mathbf{A} \sim \mathbf{I}_K$ or equivalently, $\mathbf{A} \in \mathcal{W}(K)$.

Consequently, it is not possible to obtain a vector $\mathbf{X} = \mathbf{AS}$ with independent components if matrix $\mathbf{A} \in \mathcal{M}(K)$ is not in $\mathcal{W}(K)$.

The following corollary results from Theorem 3 with $\mathbf{A} \leftarrow \mathbf{BA}$, $\mathbf{X} \leftarrow \mathbf{Y}$ and Definition 5.

Corollary 2 (Identifiability) Assume that there exists $\mathbf{A} \in \mathcal{M}(K)$ and a vector $\mathbf{S} \in \mathbb{R}^K$ of sources such that $\mathbf{X} = \mathbf{AS}$ and $\wedge_{i=3}^6 \mathcal{A}_i$ hold true. Then, the components of $\mathbf{Y} = \mathbf{BX}$ are pairwise independent if and only if $\mathbf{BA} \in \mathcal{W}(K)$ or equivalently, if and only if $\mathbf{B} \sim \mathbf{A}^{-1}$.

This result basically states that the only way to make independent the components Y_1, \dots, Y_K of $\mathbf{Y} = \mathbf{BAS}$ where \mathbf{S} satisfies $\wedge_{i=3}^6 \mathcal{A}_i$ is to have $\mathbf{B} \sim \mathbf{A}^{-1}$.

The above *identifiability theorem* links, rigorously, the BSS and ICA problems. The K sources can be recovered by finding a linear transformation \mathbf{B} such that the components of $\mathbf{Y} = \mathbf{BX}$ are independent; in this case, each output (independent component) is proportional to a distinct source signal. Formally, if the outputs are pairwise independent:

$$\forall i \in \{1, \dots, K\} \exists j(i) \in \{1, \dots, K\} : Y_i \propto S_{j(i)} ,$$

where $\cup_i \{j(i)\}$ forms a permutation of $\{1, \dots, K\}$.

Then, BSS cannot be uniquely determined : the demixing matrix $\mathbf{B} = \mathbf{A}^{-1}$ cannot be explicitly recovered. It can only be found up to the product of a diagonal (scaling) and permutation (ordering) matrix; i.e. up to a monomial transformation. Consequently, neither the order, nor the scale of the sources can be estimated via ICA (the sources can then be assumed to be unit-variance). From the ICA point of view, this is because independence is neither sensitive to the order nor to the scale of the variables. From the BSS viewpoint, this is because the mixture model $\mathbf{X} = \mathbf{AS}$ remains unchanged when i) S_j is divided by a scale factor provided that the j -th column of \mathbf{A} is scaled by the same factor, and ii) when S_i and S_j are swapped provided that the i -th and j -th columns of \mathbf{A} are also swapped. An additional source assumption will then be considered in the following:

\mathcal{A}_7 The sources are unit variance: $\text{Var}[S_i] = 1$ for all $i \in \{1, \dots, K\}$.

Combined to \mathcal{A}_5 , we have $\text{Cov}[\mathbf{S}] = \mathbf{I}_K$.

1.3 INDEPENDENCE MEASURES

From the above section, it seems that we have to express what is meant by *maximizing independence between the components of a vector*. In other words, we are thus led to find a suitable independence measure to tackle the BSS problem. A definition of independence measure is proposed below.

Definition 9 (Independence measure) *An independence measure is any mapping from a random vector \mathbf{Y} to \mathbb{R} whose maximum value is reached if and only if the components \mathbf{Y}_i of \mathbf{Y} are independent.*

Property 1 (Divergence measure and independence) *According to Def. 9, the opposite of any divergence measure of the same form as defined in \mathcal{A}_6 can be used as an independence measure, since*

$$-D \left(p_{\mathbf{Y}}(\mathbf{Y}) \middle\| \prod_{i=1}^K p_{\mathbf{Y}_i}(\mathbf{Y}_i) \right) \leq 0 , \quad (1.29)$$

with equality if and only if the joint pdf is separable into the product of the marginal densities.

According to Corollary 2, if $\mathbf{X} = \mathbf{AS}$ and $\mathbf{Y} = \mathbf{BX}$, the independence measure reaches its global maximum point when $\mathbf{B} \sim \mathbf{A}^{-1}$ or equivalently from Corollary 1, if $\mathbf{BA} \in \mathcal{W}(K)$.

Since independence measures can be obtained through a kind of “distance” between the joint density and the product of the marginal densities, let us turn to divergence measures between density functions.

1.3.1 Divergence measures between densities

Actually, we do not necessarily need a distance, in the sense that the measure is not constrained to fulfill the triangular inequality nor to be symmetric.

A non-exhaustive but rather extended list of such divergence measures can be found in [Basseville, 1989]. The most used distances in signal processing and pattern recognition are probably the f -divergences, which forms a class of “distances” independently derived in [Csiszar, 1967] and [Ali and Silvey, 1966]. This specific class of divergence measures between densities including the Bhattacharyya, Chernoff, Variational, Hellinger and Kullback-Leibler (KL) measures, see [Basseville, 1989]) is of the form:

$$\langle p \| q \rangle = f(E_p[c(L(X))]) , \quad (1.30)$$

where $f(\cdot)$ is a non-decreasing function, $E_p[\cdot]$ is the expectation with respect to p , $c(\cdot)$ is a convex function and $L(\cdot)$ is the likelihood ratio $p(\cdot)/q(\cdot)$.

Obviously, a lot of other classes of measures can be found (see e.g. [Gray et al., 1975],[Poor, 1980]), but the Ali & Silvey class and, in particular, the KL measure has been preferred in the ICA community. It is a very kind measure, in the sense that it benefits from interesting computational [Kullback, 1959, Cover and Thomas, 1991] and geometrical properties (see e.g. [Johnson and Sinanovic, 2001] and reference therein for relationship to optimal classification rates and associated manifolds and [Cardoso, 2003, 2000] for an interpretation in the ICA framework). All divergence measures belonging to the Ali & Silvey's class enjoy specific properties [Ali and Silvey, 1966]. They are not all given here because some of them would require a detailed discussion to be well understood, but we point the two following ones:

- The $\langle p\|q \rangle$ coefficient is defined for all pairs of measures p and q on the same sample space (i.e. for all pairs of densities defined on a same support Ω);
- $\langle p\|q \rangle$ is minimum when $p = q$ almost everywhere and maximum for $p \perp q$ (i.e. $\langle p\|q \rangle$ must increase when p moves apart from q).

Setting $c(x) = x \log x$ and $f(x) = x$ in Eq. (1.30), we obtain the well-known Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951, Kullback, 1959].

Definition 10 (Kullback-Leibler divergence) *Let p, q be two density functions, integrable with respect to the Lebesgue measure, and p absolutely continuous with respect to q ($\Omega(p) \subseteq \Omega(q)$). Then, the KL divergence between p, q is defined as:*

$$\text{KL}[p\|q] = \int p(x) \log \frac{p(x)}{q(x)} dx = E_p \left[\log \left(\frac{p}{q} \right) \right]. \quad (1.31)$$

1.3.1.1 KL properties

The KL obviously benefits from the “reasonable” properties of all the divergence measures of the general Ali & Silvey class; further, other specific characteristics can be emphasized [Cover and Thomas, 1991].

Proposition 1 (KL properties) *For all densities p_1, p_2, p_3 for which the quantities $\text{KL}[p_1\|p_2]$, $\text{KL}[p_1\|p_3]$ and $\text{KL}[p_2\|p_3]$ are well-defined:*

- $\text{KL}[p_1\|p_2] \geq 0$ with equality if and only if $p_1 = p_2$ almost everywhere (this results directly from Jensen's inequality);
- it is invariant under any linear invertible transformation $\varphi : \text{KL}[p_1\|p_2] = \text{KL}[\varphi(p_1)\|\varphi(p_2)]$;
- it is not a metric distance because it is not necessarily symmetric (in general: $\text{KL}[p_1\|p_2] \neq \text{KL}[p_2\|p_1]$) and it usually violates the triangular inequality ($\text{KL}[p_1\|p_2] + \text{KL}[p_2\|p_3] \not\geq \text{KL}[p_1\|p_3]$);

Other divergence measures between densities can be found. For instance, we could take one of the following symmetric quantities

$$\langle p\|q\rangle = \int |p(x) - q(x)|dx , \quad (1.32)$$

or

$$\langle p\|q\rangle = \int (p(x) - q(x))^2 dx . \quad (1.33)$$

However, the benefits gained from considering the KL divergence in BSS are so considerable from the computational simplicity viewpoint that no other divergence measure has been seriously investigated, to our knowledge, in the BSS framework, even though they can be used, exactly as the KL, to derive an independence measure.

1.3.1.2 From KL to mutual information

From Property 1, the KL between p_Y (the joint pdf of the multivariate random vector) and $\prod_{i=1}^K p_{Y_i}$ (the product of the marginal pdf of the components of Y) seems to be an interesting independence measure; it is called the *mutual information*.

Definition 11 (Mutual information) *The mutual information of a random vector $Y = [Y_1, \dots, Y_K]^T$ is defined as*

$$\text{KL}\left[p_Y \middle\| \prod_{i=1}^K p_{Y_i}\right] . \quad (1.34)$$

This divergence measure is also equivalently noted $\text{KL}(Y)$.

Proposition 2 (Mutual information properties) *The mutual information properties result from the KL ones but, further remarkable results are the following.*

- $\text{KL}(Y) \geq 0$ with equality if and only if the components of Y are mutually independent; hence, $-\text{KL}(Y)$ is an independence measure.
- It is symmetric: $\text{KL}(Y) = \text{KL}[p_Y, \prod_{i=1}^K p_{Y_i}] = \text{KL}[\prod_{i=1}^K p_{Y_i}, p_Y]$
- Let us define $\varphi(Y) = [\varphi_1(Y_1), \dots, \varphi_K(Y_K)]$ where the K φ_i 's are linear invertible with existing derivatives and derivable inverses mappings (i.e. diffeomorphisms). Then $\text{KL}(Y) = \text{KL}(\varphi(Y))$.

1.3.2 Other measures of independence

The independence can also be measured by some means different from a divergence between the joint density of a random vector and the product of the marginal densities of its components. Remind that K non-Gaussian rv $\mathbf{Y}_1, \dots, \mathbf{Y}_K$ are independent if all their cross-cumulants vanish. Hence, a measure of independence would be a functional of positive mappings of all the cross-cumulants; independence would then be reached if and only the functional vanishes. In practice however, this is not feasible: we have to consider only a finite number of cross-cumulants. Most often, existing methods assume a whitening pre-processing (see Section 1.5) so that the outputs are decorrelated, and they only cancel a finite number of cross-cumulants. For instance, the criterion

$$-\sum_{ijkl \neq iiii} \text{cum}_{ijkl}^2(\mathbf{Y}) \quad (1.35)$$

can be seen as an approximate measure of independence between the \mathbf{Y}_i if $E[\mathbf{Y}_i \mathbf{Y}_j] = E[\mathbf{Y}_i]E[\mathbf{Y}_j]$, $i \neq j$ (the fourth-order cumulant was defined in Eq. 1.28, p. 12); but this approximation suffices to solve the BSS problem [Comon, 1994]. Minimizing the last criterion is equivalent to maximizing $\sum_i \text{cum}_{iiii}^2(\mathbf{Y})$ [Comon, 1994].

A variant of this criterion,

$$-\sum_{ijkl \neq ijkk} \text{cum}_{ijkl}^2(\mathbf{Y}) \quad (1.36)$$

has also been proposed [Cardoso, 1998, Cardoso and Souloumiac, 1993].

Similarly, another approximate measure would be to compute a linear combination of positive mappings of

$$E[f(\mathbf{Y}_i)g(\mathbf{Y}_j)] - E[f(\mathbf{Y}_i)]E[g(\mathbf{Y}_j)] \text{ for } i, j \in \{1, \dots, K\}, i \neq j \quad (1.37)$$

for a finite number of functions f, g . The possibly non-linear functions $f, g \in \mathbb{F}$ capture the higher-order information on \mathbf{X}, \mathbf{Y} , not only their covariance. The pioneering solution to the BSS problem, which was proposed by Hérault and Jutten in 1991, was based on this approach; they used $f(y) = y^3$ and $g(y) = \arctan(y)$ [Jutten and Hérault, 1991].

This is the method used in [Bach and Jordan, 2002]: independence is reached if and only if the following generalized non-linear correlation coefficient is zero:

$$\rho_{\mathbb{F}} = \max_{f, g \in \mathbb{F}} \text{Corr}[f(\mathbf{X}), g(\mathbf{Y})] = \max_{f, g \in \mathbb{F}} \frac{\text{Cov}[f(\mathbf{X}), g(\mathbf{Y})]}{\sqrt{\text{Var}[f(\mathbf{X})]\text{Var}[g(\mathbf{Y})]}} . \quad (1.38)$$

A review of independence measures for BSS can be found in [Achard, 2003].

1.4 EXTRACTION SCHEMES AND CONTRAST FUNCTION DEFINITION

In Section 1.2.3, a relation between independence (and thus ICA) and BSS was pointed out by Theorem 3 (p. 13): maximizing any independence measure (as defined in Def. 9, p. 14) of \mathbf{BX} will provide a demixing matrix being PD-equivalent to \mathbf{A}^{-1} , i.e. such that each of the output \mathbf{Y}_i will be proportional to a distinct source \mathbf{S}_j . However, even under the usual assumption on the independence of the sources, source separation cannot be seen, generally speaking, to be equivalent to independent component analysis; this is e.g. the case where only a subset of whatever sources is needed.

Independence measures are rather restrictive criteria; they implicitly require that all the outputs be considered at the same time. For instance, if one desires to extract a single source, the viewpoint of independence maximization cannot be easily adopted as a single signal is considered, while independence is a relative quantity. In a more general framework, one may desire to separate P signals from K mixtures. In this case, ICA is not equivalent to BSS, as shown in the next example.

Example 2 (ICA is not BSS) Let $\mathbf{W} = \mathbf{BA}$, $\mathbf{Y}_j = \sum_{i=1}^K W_{ji} \mathbf{S}_i$, $j \in \{1, \dots, P\}$ where W_{ij} is the (i,j) -th entry of \mathbf{W} . Assume $P = 2$ for simplicity. We know from the Theorem 2 (the converse form of the Darmois-Skitovitch theorem, p. 12) that if $W_{1i}W_{2i} = 0$ for all $1 \leq i \leq K$, then \mathbf{Y}_1 is independent from \mathbf{Y}_2 .

In other words, if one can find $\mathbf{B} \in \mathcal{M}^{P \times K}$ such that \mathbf{BA} has exactly one non-zero element per column, the entries of \mathbf{BX} are independent provided that at most one source is Gaussian. The problem is that such matrices are not necessarily subPD-equivalent to \mathbf{A}^{-1} , that is, we can find $\mathbf{B} \in \mathcal{M}^{P \times K}$ such that the components of \mathbf{BX} are independent but $\mathbf{BA} \notin \mathcal{W}^{P \times K}$. For instance setting $P = 2$ and $K = 4$, a demixing matrix \mathbf{B} such that

$$\mathbf{BA} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad (1.39)$$

yields independent outputs, but none of them is proportional to a source signal: $\mathbf{BA} \notin \mathcal{W}^{P \times K}$.

From the above example, we conclude that ICA only solves specific schemes of the BSS problem, in which the number of outputs equals the number of sources involved in the mixture. Therefore, in order to deal with separation schemes more general than the simultaneous separation of all the K sources, it is necessary to find BSS criteria that are no more necessarily “pure” independence measure.

The purpose of this section is to review three extraction schemes that can be used to recover (part of) the demixing matrix. For each of these schemes,

the BSS problem is rewritten as an optimization problem; based on the contrast function definition, solving the BSS problem reduces to maximizing a criterion, if the latter belongs to the specific class of the BSS contrast functions.

1.4.1 Extraction schemes

When recovering the sources, three approaches can be adopted to estimate \mathbf{B} such that $\mathbf{B} \sim_u \mathbf{A}^{-1}$. First, the demixing matrix \mathbf{B} can be estimated globally, by maximizing a *simultaneous* contrast function. Second, $P \leq K$ rows \mathbf{b}_j of \mathbf{B} can be estimated one by one, yielding the sources sequentially. Third, $P \leq K$ rows of \mathbf{B} can be estimated simultaneously. These three approaches are respectively named *simultaneous*, *deflation* and *partial* separation.

1.4.1.1 Simultaneous separation

In simultaneous extraction, one is led to maximize a criterion $f(\mathbf{B})$ with respect to a $K \times K$ matrix \mathbf{B} , so that the rows of \mathbf{B} are estimated all at once. After convergence, we shall have $\mathbf{B} \sim \mathbf{A}^{-1}$.

1.4.1.2 Deflation separation

An alternative method is to compute the rows of \mathbf{B} one by one. Instead of maximizing the function $f(\mathbf{B})$ directly in a $K \times K$ -dimensional space, K “sub-functions” are estimated iteratively. The sources are recovered by sequentially maximizing $f(\mathbf{b}_1), \dots, f(\mathbf{b}_K)$, where the \mathbf{b}_i are the rows of the target matrix \mathbf{B} . A decorrelation constraint is added in order to avoid recovering twice a same source: at each step, the i -th estimated source must be uncorrelated to the $i - 1$ previously extracted sources.

The deflation method has an advantage compared to the simultaneous approach: it allows one to extract $P \leq K$ sources by maximizing sequentially $f(\mathbf{b}_1), \dots, f(\mathbf{b}_P)$, and then considering $\mathbf{b}_1\mathbf{X}, \dots, \mathbf{b}_P\mathbf{X}$ as the P estimates. However, this sequential technique may suffer from the cumulation of errors resulting from the decorrelation constraint imposed between the rows.

1.4.1.3 Partial separation

Recently, a new kind of extraction has been introduced in [Pham, 2006b]: it consists in simultaneously extracting P among K sources, for any $P \leq K$. This method yields all the K sources if $P = K$ (and thus reduces to the simultaneous separation scheme) and yields the first output just as the deflation method if $P = 1$.

1.4.2 Contrast functions

In this section, the BSS problem is formalized as an optimization problem for each of the extraction schemes. For each scheme, we define the concept of *contrast function*, which is the objective of the optimization problem.

1.4.2.1 Simultaneous separation

Definition 12 (Simultaneous BSS (S-BSS) contrast) A simultaneous BSS (S-BSS) contrast is a mapping $\mathcal{C}(\cdot) : \mathcal{M}(K) \mapsto \mathbb{R}$ being invariant under PD-equivalency preserving transforms and satisfying

$$\operatorname{argmax}_{\mathbf{B} \in \mathcal{M}(K)} \mathcal{C}(\mathbf{B}) = \{\mathbf{B} \in \mathcal{M}(K) : \mathbf{B} \sim \mathbf{A}^{-1}\} , \quad (1.40)$$

or equivalently:

$$\operatorname{argmax}_{\mathbf{B} \in \mathcal{M}(K)} \mathcal{C}(\mathbf{B}) = \{\mathbf{B} \in \mathcal{M}(K) : \mathbf{B}\mathbf{A} \in \mathcal{W}(K)\} . \quad (1.41)$$

In the above definition, the mathematical expression $\operatorname{argmax}_{x \in \mathcal{X}} f(x)$ has to be understood as the set of points \mathcal{X}^* in $\operatorname{dom} f \cap \mathcal{X}$ such that the function f reaches its global maximum value over the set $\operatorname{dom} f \cap \mathcal{X}$ at, and only at points in \mathcal{X}^* .

Note that this definition differs slightly from the *contrast* definition first given in [Comon, 1994]. However, we define here a *BSS contrast*, i.e. a contrast in the framework of BSS; this corresponds now to the usual common meaning of a contrast for BSS, as accepted by the related community (see e.g. [Cardoso, 1998], etc). The major differences are the following. First, it is no more supposed that a contrast is a mapping *from the set of densities*. It is here understood as a mapping from $\mathcal{M}(K)$. Second, the *only if* statement in the third property was not required in Comon's definition. This additional requirement yielded, with Comon's terminology, to a *discriminating* contrast. However, this terminology will be used for another property, which will be stated in Chapter 3.

From the Identifiability theorem (Corollary 2) and the independence measure definition, the following corollary trivially holds.

Corollary 3 Under assumptions $\wedge_{i=1}^6 \mathcal{A}_i$, any independence measure being simultaneously scale-invariant and permutation-invariant is a simultaneous contrast function.

1.4.2.2 Deflation separation

Definition 13 (Deflation BSS contrast) A mapping $\mathcal{C}(\cdot) : \mathbb{R}^K \mapsto \mathbb{R}$ is a deflation BSS (D-BSS) contrast if it fulfills the following conditions:

- it is scale-invariant: $\mathcal{C}(\mathbf{b}) = \mathcal{C}(\alpha\mathbf{b})$ for any $\alpha \in \mathbb{R}_0$;

- the global maximum point is attained when $\mathbf{b}\mathbf{A}$ is proportional to a row of \mathbf{I}_K :

$$\underset{\mathbf{b} \in \mathbb{R}^K}{\operatorname{argmax}} \mathcal{C}(\mathbf{b}) \subseteq \{\mathbf{b} \in \mathbb{R}^K : \mathbf{b}\mathbf{A} \propto \mathbf{e}_i, i \in \{1, \dots, K\}\} , \quad (1.42)$$

or equivalently, by noting $\|\mathbf{x}\|$ the Euclidean norm of \mathbf{x} :

$$\underset{\mathbf{b} \in \mathbb{R}^K}{\operatorname{argmax}} \mathcal{C}(\mathbf{b}) \subseteq \{\mathbf{b} \in \mathbb{R}^K : \|\mathbf{b}\mathbf{A}\| = \|\mathbf{b}\mathbf{A}\|_\infty\} ; \quad (1.43)$$

- the criterion allows complete extraction. Let \mathcal{S}_k be the set of the source indices that have been extracted at the k -th step: for any $1 \leq i \leq k$, $\mathbf{b}_i \mathbf{X} \propto \mathcal{S}_j$, $j \in \mathcal{S}_k$. Then, under the constraint that $\mathbf{b}\mathbf{b}_i^T = 0$ for $i \in \{1, \dots, K\} \setminus \mathcal{S}_k$:

$$\underset{\mathbf{b} \in \mathbb{R}^K}{\operatorname{argmax}} \mathcal{C}(\mathbf{b}) \subseteq \{\mathbf{b} \in \mathbb{R}^K : \mathbf{b}\mathbf{A} \propto \mathbf{e}_i, i \in \{1, \dots, K\} \setminus \mathcal{S}_k\} . \quad (1.44)$$

1.4.2.3 Partial separation

Definition 14 (Partial BSS contrast) A mapping $\mathcal{C}(\cdot) : \mathcal{M}^{P \times K} \mapsto \mathbb{R}$ is a partial BSS (P -BSS) contrast if it fulfills the following conditions:

- it is scale-invariant: $\mathcal{C}(\mathbf{B}) = \mathcal{C}(\Lambda \mathbf{B})$ for any matrix $\Lambda \in \mathcal{D}(P)$;
- it is order-invariant: $\mathcal{C}(\mathbf{B}) = \mathcal{C}(\Pi \mathbf{B})$ for any matrix $\Pi \in \mathcal{P}(P)$;
- the set of the global maximum points are $P \times K$ non-mixing matrices:

$$\underset{\mathbf{B} \in \mathcal{M}^{P \times K}}{\operatorname{argmax}} \mathcal{C}(\mathbf{B}) \subseteq \{\mathbf{B} \in \mathcal{M}^{P \times K} : \mathbf{B} \sim_u \mathbf{A}^{-1}\} , \quad (1.45)$$

or equivalently,

$$\underset{\mathbf{B} \in \mathcal{M}^{P \times K}}{\operatorname{argmax}} \mathcal{C}(\mathbf{B}) \subseteq \{\mathbf{B} \in \mathcal{M}^{P \times K} : \mathbf{B}\mathbf{A} \in \mathcal{W}^{P \times K}\} . \quad (1.46)$$

Note that the first two items are equivalent to require $\mathcal{C}(\mathbf{B}) = \mathcal{C}(\mathbf{MB})$ if $\mathbf{M} \sim \mathbf{I}_P$.

It is a kind of compromise between deflation and simultaneous approaches, combining the advantages of both separation schemes, in the sense that we can limit the computational load if only $P < K$ sources are needed and on the other hand, the cumulation of errors resulting from the orthogonalization constraint is avoided³.

³We would like to point out here that generally speaking, there is no information about which subset of P sources will be extracted.

1.5 WHITENING PREPROCESSING AND GEODESIC SEARCH

Whitening is a usual preprocessing to BSS, either for simultaneous or deflation extraction schemes. It does *half the work* of ICA, in the sense that the dimensionality of the BSS problem is approximatively divided by a factor 2 thanks to the whitening. In other words, *half the job* is done by using an algebraic technique (reminded in Section 1.5.1), so that *orthogonal BSS contrast functions* can be proposed (Section 1.5.2). In the remaining adaptive optimization step, the argument space is limited in such a way that the possible adaptive search will be managed in a lower dimensional subspace of the original space of the demixing matrices (Section 1.5.4).

1.5.1 Whitening

Whitening a data vector \mathbf{X} consists in jointly i) centering the data (unnecessary if \mathcal{A}_4 holds), ii) linearly transform them in such a way that they become uncorrelated ($\text{Cov}[\mathbf{X}]$ is diagonal), and iii) scaling the \mathbf{X}_i so that they become unit-variance (yielding $\text{Cov}[\mathbf{X}] = \mathbf{I}_K$).

Definition 15 (whitening matrix) A matrix $\mathbf{V} \in \mathcal{M}(K)$ is a *whitening matrix* of a random vector $\mathbf{X} \in \mathbb{R}^K$ if $\mathbf{Z} = \mathbf{V}\mathbf{X}$ is a white random vector: $E[\mathbf{Z}] = \mathbf{0}$ and $E[\mathbf{Z}\mathbf{Z}^T] = \mathbf{I}_K$.

A whitening matrix \mathbf{V} of \mathbf{X} can be found via eigenvalue decomposition (EVD).⁴ Let us denote by $\boldsymbol{\Lambda}$ a diagonal matrix whose diagonal entries are the eigenvalues of the $K \times K$ covariance matrix $\text{Cov}[\mathbf{X}]$, and let \mathbf{U} be an orthogonal matrix (because $\text{Cov}[\mathbf{X}] = \text{Cov}^T[\mathbf{X}]$) whose columns are unit-norm eigenvectors such that $\mathbf{e}_i \mathbf{U}^T$ is the eigenvector associated to the eigenvalue located at the i -th diagonal element of $\boldsymbol{\Lambda}$. Then,

$$\mathbf{V} \doteq \boldsymbol{\Lambda}^{-1/2} \mathbf{U}^T \quad (1.47)$$

is a whitening matrix. This results from the following equalities:

$$\begin{aligned} E[\mathbf{V}\mathbf{X}(\mathbf{V}\mathbf{X})^T] &= \mathbf{V} \underbrace{E[\mathbf{X}\mathbf{X}^T]}_{=\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T} \mathbf{V}^T \\ &= \boldsymbol{\Lambda}^{-1/2} \underbrace{\mathbf{U}^T}_{\mathbf{I}_K} \underbrace{\mathbf{U}}_{\mathbf{I}_K} \boldsymbol{\Lambda} \underbrace{\mathbf{U}^T}_{\mathbf{I}_K} \underbrace{\mathbf{U}}_{\mathbf{I}_K} \boldsymbol{\Lambda}^{-1/2} \end{aligned} \quad (1.48)$$

$$= \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\Lambda} \boldsymbol{\Lambda}^{-1/2} \quad (1.49)$$

$$= \mathbf{I}_K . \quad (1.50)$$

$$= \mathbf{I}_K . \quad (1.51)$$

⁴Singular value decomposition (SVD) can also be used but is not considered here, except briefly in the last Chapter.

Observe that in the above developments, it is assumed that all the eigenvalues are non-zero in order that $\Lambda^{-1/2}$ exist since $\det \Lambda = \prod_{i=1}^K \Lambda(i, i)$ (in other words, $\Lambda \in \mathcal{D}(K) \subset \mathcal{M}(K)$). Clearly, this is the case since $\text{Cov}[\mathbf{X}]$ is symmetric (the eigenvalues are real) and because the covariance matrix is positive semi-definite (all the eigenvalues will be non-negative) leading to $\det \text{Cov}[\mathbf{X}] \geq 0$. Further, $\text{Cov}[\mathbf{X}] \in \mathcal{M}(K)$ since $\text{Cov}[\mathbf{X}]$ is full-rank:

$$\text{rank}(\text{Cov}[\mathbf{X}]) = \text{rank}(\mathbf{A} \text{Cov}[\mathbf{S}] \mathbf{A}^T) \stackrel{A_7}{=} \text{rank}(\mathbf{A} \mathbf{A}^T) = \text{rank}(\mathbf{A}) = K . \quad (1.52)$$

Consequently, $\det \text{Cov}[\mathbf{X}] > 0$ and, on the other hand $\det \text{Cov}[\mathbf{X}] = \det \Lambda \det^2 \mathbf{U} = \det \Lambda$, implying $\Lambda \in \mathcal{D}(K) \subset \mathcal{M}(K)$.

An important subset of $\mathbb{R}^{K \times K}$ is the orthogonal group.

Definition 16 (Orthogonal group) *The orthogonal group of degree K is the subset of orthogonal matrices :*

$$\mathcal{O}(K) \doteq \{\mathbf{M} \in \mathcal{M}(K) : \mathbf{M} \mathbf{M}^T = \mathbf{I}_K\} . \quad (1.53)$$

A specific subset of this group is the special orthogonal group:

Definition 17 (Special orthogonal group) *The special orthogonal group is the subset of $\mathcal{O}(K)$ corresponding to rotation matrices, i.e.*

$$\mathcal{SO}(K) \doteq \{\mathbf{M} \in \mathcal{O}(K) : \det \mathbf{M} = +1\} . \quad (1.54)$$

The following property can be easily proved.

Property 2 (Whiteness preservation under orthogonal transform) *Let \mathbf{V} be a whitening matrix of $\mathbf{X} \in \mathbb{R}^K$. Then, for any orthogonal matrix $\mathbf{R} \in \mathcal{O}(K)$, \mathbf{RV} is a whitening matrix of \mathbf{X} .*

The above property states that whiteness property is preserved under orthogonal transforms and, consequently, that the whitening matrix of a random vector is not unique. It results directly from the fact that $E[(\mathbf{RV}\mathbf{X})(\mathbf{RV}\mathbf{X})^T] = E[\mathbf{R}\mathbf{Z}\mathbf{Z}^T\mathbf{R}^T]$ where $\mathbf{Z} \doteq \mathbf{V}\mathbf{X}$. The last expectation reduces to $\mathbf{R}\mathbf{R}^T = \mathbf{I}_K$ since \mathbf{Z} is a white random vector.

The whitening preprocessing is seen as solving *half of the ICA problem*. Indeed, even if not known, \mathbf{VA} reduces to an orthogonal matrix, because:

$$E[\mathbf{ZZ}^T] = E[\mathbf{VAS}(\mathbf{VAS})^T] \quad (1.55)$$

$$= \mathbf{VAE}[\mathbf{SS}^T]\mathbf{A}^T\mathbf{V}^T \quad (1.56)$$

$$\stackrel{A_7}{=} (\mathbf{VA})(\mathbf{VA})^T \quad (1.57)$$

$$= \mathbf{I}_K , \quad (1.58)$$

where the last equality results from the whiteness of the zero-mean vector \mathbf{Z} .

Because of the group structure of $\mathcal{O}(K)$, the inverse of the transfer matrix \mathbf{VA} from \mathbf{S} to \mathbf{Z} is also included in $\mathcal{O}(K)$. Consequently, one can restrict the search of a demixing matrix \mathbf{B} from $\mathcal{M}(K)$ to $\mathcal{O}(K)$ and even to $\mathcal{SO}(K)$, since assuming $\det \mathbf{B} = +1$ does not add any indeterminacy on the recovered sources, as shown by the following lemma.

Lemma 4 *For any matrix $\mathbf{M}_1 \in \mathcal{O}(K)$, there exists $\mathbf{M}_2 \in \mathcal{SO}(K)$ such that $\mathbf{M}_2 \sim \mathbf{M}_1$.*

If $\mathbf{M}_1 \in \mathcal{O}(K)$, $|\det \mathbf{M}_1| = 1$. Hence, by definition, $\mathbf{M}_1 \in \mathcal{SO}(K)$ and we can take $\mathbf{M}_2 = \mathbf{M}_1$. Assume now that $\det \mathbf{M}_1 = -1$. Noting by $\mathbf{I}_K^{(-i)}$ the identity matrix \mathbf{I}_K in which the i -th diagonal element is replaced by its opposite. Then, $\det \mathbf{I}_K^{(-i)} \mathbf{M}_1 = \det \mathbf{I}_K^{(-i)} \det \mathbf{M}_1 = -\det \mathbf{M}_1 = 1$. But, by definition of the PD-equivalency, $\mathbf{M}_2 = \mathbf{I}_K^{(-i)} \mathbf{M}_1$ is PD-equivalent to \mathbf{M}_1 and $\mathbf{M}_2 \in \mathcal{SO}(K)$.

□

Then, since the dimension of $\mathcal{SO}(K)$ is $K \times (K-1)/2$, the number of elements to be estimated in the demixing model is approximatively divided by a factor two. The whitening preprocessing reduces the dimensionality of the problem; an orthogonal contrast can then be used. An *orthogonal contrast* is a contrast whose argument is constrained to be in $\mathcal{SO}(K)$.

1.5.2 Orthogonal contrast functions

Definition 18 (Orthogonal BSS contrast) *An orthogonal simultaneous (resp. deflation) BSS contrast is a simultaneous (resp. deflation) BSS contrast where the mixing matrix is assumed to be a rotation matrix ($\mathbf{A} \in \mathcal{SO}(K)$) and the demixing matrix is always constrained to be in the special orthogonal group. A partial orthogonal BSS contrast is a partial contrast whose argument is constrained to be a semi-orthogonal matrix, that is where $\mathbf{BB}^T \in \mathcal{SO}(K)$. The orthogonal BSS contrasts are noted \mathcal{C}^\perp .*

Obviously, the set of orthogonal contrasts is included in the set of (global) BSS contrasts.

Remark 1 *Note that whatever the mixing matrix \mathbf{A} , one can still use an orthogonal contrast. In order to do that, one can deal with a whitened version of the mixtures since as shown in Section 1.5.1: $\mathbf{VA} \in \mathcal{O}(K)$ if \mathbf{V} is a whitening matrix of $\mathbf{X} = \mathbf{AS}$.*

Assume $\mathbf{X} \leftarrow \mathbf{VX}$. Then, if $\mathbf{B}^ \in \mathcal{SO}(K)$ maximizes the orthogonal contrast, the demixing matrix satisfies $\mathbf{B}^* \sim (\mathbf{VA})^{-1}$ and from Lemma 2 (p. 9), $\mathbf{B}^* \mathbf{V} \sim \mathbf{A}^{-1}$.*

Since the mixing matrix is not necessarily orthogonal, we shall always consider $\mathbf{X} \leftarrow \mathbf{VX}$ and $\mathbf{A} \leftarrow \mathbf{VA}$ when dealing with orthogonal contrast functions.

1.5.3 Angular parametrization in the K=2 case

It has been explained that if \mathbf{X} is whitened, \mathbf{A} can be seen to be an orthogonal matrix. In two dimensions, an orthogonal matrix is fully determined by a single angle, called here the *mixing angle* ϕ . As explained above, one can freely assume that \mathbf{A} is a pure rotation matrix ($\det \mathbf{A} = +1$). Hence, the mixing and prewhitening steps can be expressed as follows:

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} = \begin{bmatrix} \sin \phi & \cos \phi \\ -\cos \phi & \sin \phi \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix}. \quad (1.59)$$

Furthermore, under the additional $E[\mathbf{Y}\mathbf{Y}^T] = \mathbf{I}_K$ constraint, \mathbf{W} also reduces to an orthogonal (assumed rotation) matrix, parametrized by a single *unmixing angle* φ . Hence, for $K = 2$, the input-output model becomes:

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \sin(\phi + \varphi) & \cos(\phi + \varphi) \\ -\cos(\phi + \varphi) & \sin(\phi + \varphi) \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix}. \quad (1.60)$$

Hence, the BSS problem reduces to finding the unknown initial angle ϕ only knowing \mathbf{X} and \mathbf{Y} by adjusting φ . Let us define $\theta \doteq \phi + \varphi$. The angle ϕ is fixed, since \mathbf{A} is a constant matrix, but θ is unknown, and may vary via φ . Consequently, the transfer matrix, also noted $\mathbf{W}(\theta)$ for clarity, is non-mixing if and only if we have found *blindly* (ϕ is unknown) $\varphi = \varphi^*$ such that $\varphi^* = k\pi/2 - \phi$, $k \in \mathbb{Z}$. When a single output is considered in the $K = 2$ case, it will often be noted

$$\mathbf{Y}_\theta \doteq \sin(\theta)\mathbf{S}_1 + \cos(\theta)\mathbf{S}_2, \quad (1.61)$$

i.e. it corresponds to the first output \mathbf{Y}_1 of the model given in Eq. (1.60) with $\theta = \phi + \varphi$.

1.5.4 Manifold-constrained problem and geodesic optimization

From the above subsection, we conclude that under pre-whitening constraint, there always exists $\mathbf{B} \in \mathcal{SO}(K)$ such that $\mathbf{B} \sim (\mathbf{VA})^{-1}$. We can thus restrict the search of demixing matrix to $\mathcal{SO}(K)$. Lie groups such as e.g. $\mathbb{R}^{K \times K}$, $\mathcal{M}(K)$, $\mathcal{O}(K)$ or $\mathcal{SO}(K)$ can be given a Riemannian *manifold* structure [Amari, 1998, Plumley, 2004, Chefd'Hotel et al., 2004].

Definition 19 (Manifold) A manifold is a topological space which is locally Euclidean.

Without entering into details, various definitions of the *manifold* object exist: basically, some of them suppose that the manifold is “smooth everywhere”, the others assume that the manifold is locally flat almost everywhere. As an example, depending of the definition, the boundary of a square is a manifold or not, but in any case, it is not a “smooth manifold”, because of the corners [Absil et al., Lee, 2003].

In the sequel, the *set* notation will be used for the manifolds. Generally, a manifold will be noted \mathcal{M} (not related to the set of matrices $\mathcal{M}(K)$ and $\mathcal{M}^{P \times K}$). In other words, it is a mathematical space in which (almost) every point has a neighborhood which resembles the Euclidean space, but in which the global structure may be more complicated. When these manifolds are embedded in \mathbb{R}^K , they are named “sub-manifolds embedded in \mathbb{R}^K ”. Any open subsets in \mathbb{R}^K forms a sub-manifold. Curves and circles are examples one-dimensional (smooth) sub-manifolds embedded in \mathbb{R}^K ($K \geq 2$), also named one-manifold for short, because locally, every point has a neighborhood that resembles a line. The surface of a sphere is an example of a two-dimensional manifold because locally, the neighborhood of every points on a sphere looks like a plane (the surface of the Earth was formerly believed to be a plane because on the human scale, the surface of a sphere looks “flat”). In the following, it is always assumed that the D-manifold is embedded in \mathbb{R}^K (for some K large enough), and the prefix “D-” and “sub-” will be omitted when unnecessary. In general, any object embedded in \mathbb{R}^K which is “nearly flat” on small scales is a manifold embedded in \mathbb{R}^K provided that K is large enough (to, indeed, embed it). On a manifold, the basic usual rules of geometry do no more hold as it is not a vector space. Generally speaking, the sum of the angles of a (curved) triangle laying on a manifold does not equal π , and summing two vectors belonging to this space does not result in a third vector belonging to the manifold. By contrast, the definition says that at each point of the manifold, there exists a tangent space on which we can use the common calculus. Therefore, if the aforementioned triangle is sufficiently small, the sum approximatively equals π . The manifolds are seen here to be (possibly lower-dimensional) spaces (i.e. subspaces) embedded in a higher-dimensional Euclidean space. They may be created by a kind of “constraint”: a centered circle of radius r is a one-manifold embedded in \mathbb{R}^2 associated to the vectors in \mathbb{R}^2 having a Euclidean norm equal to r . One can also associate manifolds to the sets $\mathcal{O}(K)$ and $\mathcal{SO}(K)$; these are $K(K - 1)/2$ -dimensional manifolds embedded in the set of square matrices $\mathbb{R}^{K \times K}$. Most of the time, one deals with “smooth” manifolds (because it is implicitly required in the definition or because it is often required when dealing with other definitions); basically, they are manifolds with functional structure (e.g. parametric equations [Lee, 2003]). The unit circle in the xy-plane (defined by the constraint $x^2 + y^2 = 1$) is the smooth manifold with parametric equations ($x = \cos \theta$, $y = \sin \theta$). The parametric equations (when they exist) are friendly because they can be integrated and differentiated termwise. Informally, this means that the notion of “differentiability” exists on smooth manifolds (differential geometry is nothing else than the study of calculus on smooth manifolds). As an illustration, let $f(x, y) = xy^2$ s.t. $ax^2 + by^2 = 1$ for two scalar numbers a, b . How does this function change for a *small variation* of (x, y) ? This question seems difficult to answer because of the ambiguity about the meaning of “a small variation of (x, y) ”. The function is only defined on the set $ax^2 + by^2 = 1$ and not on \mathbb{R}^2 ; consequently, the new pair of coordinates $(x, y) + (\delta x, \delta y)$ is required to fulfill the constraint. Using the parametrization $x = \cos \theta / \sqrt{a}$, $y = \sin \theta / \sqrt{b}$, the constraint is implicitly fulfilled and it makes

sense to compute the derivative of $f(\theta) = \cos \theta \sin^2 \theta / \sqrt{ab^2}$ with respect to θ . For a small increment $\delta\theta$ of θ we have

$$f(\theta + \delta\theta) \approx f(\theta) + \frac{\sin \theta}{\sqrt{ab^2}}(2 \cos^2 \theta - \sin \theta)\delta\theta .$$

The manifold associated to the definition domain of the above function can be seen to be “smooth” because a *differentiable parametrization* is possible⁵. By contrast, the set of points satisfying $xy = 0$ does not form a smooth manifold because of the intersection.

An additional interesting property of a manifold is its connectedness. Intuitively, a manifold is connected if any pair of points can be joined by a piecewise smooth curve belonging to the manifold.

Property 3 (Properties of $\mathcal{O}(K)$ and $\mathcal{SO}(K)$) *The smooth manifolds associated to $\mathcal{O}(K)$ and $\mathcal{SO}(K)$ satisfy the following properties*

- $\mathcal{O}(K)$ is a manifold composed of $\mathcal{SO}(K)$ and $\{\mathbf{B} \in \mathcal{O}(K) : \det \mathbf{B} = -1\}$;
- $\mathcal{SO}(K)$ is a connected manifold containing the identity matrix \mathbf{I}_K ;
- the restriction of the neighborhood of a given point $\mathbf{B} \in \mathbb{R}^{K \times K}$ to the manifold induced by $\mathcal{O}(K)$ is a subset of the neighborhood of \mathbf{B} in the whole $\mathbb{R}^{K \times K}$ space (recall that $\mathcal{O}(K) \subset \mathbb{R}^{K \times K}$). This is also true for $\mathbf{B} \in \mathcal{SO}(K)$, since $\mathcal{SO}(K)$ is a connected subgroup of $\mathcal{O}(K)$.

More details about manifolds can be found in [Absil et al.].

1.6 ADAPTIVE MAXIMIZATION OF CONTRAST FUNCTIONS

In Section 1.4.2 contrast functions have been defined in order to rewrite the BSS problem as an optimization problem. However, the maximization methods have not been discussed yet. A lot of optimization techniques exist. They can be based on algebraic or adaptive methods.

Example 3 (Algebraic vs Adaptive solution) Let $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_K]^T$. The matrix $\mathbf{V} = \Lambda^{-1/2} \mathbf{U}^T$ (see Eq. (1.47)) is an algebraic solution to the problem

$$\operatorname{argmax}_{\mathbf{M} \in \mathcal{M}(K)} \left\{ - \sum_{i \neq j} \left| [\operatorname{Cov}[\mathbf{MX}]]_{ij} \right| \right\} , \quad (1.62)$$

⁵The simple approach saying that \mathcal{M} is a differentiable manifold is equivalent to the parametric equations of \mathcal{M} are differentiable has not been proved to exactly correspond to the formal definition of “manifold differentiability” (that can be of class C^k , like for functions). However, this intuitive explanation generally holds true (at least in the sense “differentiable parametric equations” \Rightarrow “differentiable manifold”) and anyway, this intuitive way of seeing things suffice for our purposes

where $[\mathbf{M}]_{ij} = M_{ij}$ is the (i, j) -th entry of \mathbf{M} , because the maximum value of this expression is zero and $\text{Cov}[\mathbf{MX}]$ is diagonal if \mathbf{M} is a whitening matrix of \mathbf{X} (i.e. if $\mathbf{M} = \mathbf{UV}$ with $\mathbf{U} \in \mathcal{O}(K)$ and \mathbf{V} given by Eq. (1.47)). As a matter of fact, the solution is provided by an eigenvalue decomposition of $\text{Cov}[\mathbf{X}]$. The result can be expressed as a closed form expression. The problem of finding the maximum reduces to the problem of finding the eigenvalues and the associated eigenvectors.

An adaptive solution to the problem would consist in choosing an initial point in $\mathcal{M}(K)$, say $\mathbf{M}^{(0)}$ and then modifying $\mathbf{M}^{(t+1)} \leftarrow \mathbf{M}^{(t)} + \Delta(\mathbf{M}^{(t)})$ where $\Delta(\mathbf{M}^{(t)})$ is such that

$$-\sum_{i \neq j} \left| \left[\text{Cov}[\mathbf{M}^{(t)} \mathbf{X}] \right]_{ij} \right| \leq -\sum_{i \neq j} \left| \left[\text{Cov}[\mathbf{M}^{(t+1)} \mathbf{X}] \right]_{ij} \right|. \quad (1.63)$$

The major difference between algebraic and adaptive optimization techniques is not that the former is non-iterative while the latter is, because iterative schemes can be required for computing (estimating) the parameters of the closed form solution. As an illustration, the computation of the eigenvalues and eigenvectors in Ex. 3 (p. 27) may require an iterative scheme (e.g. iteration of QR decompositions). The main difference is rather that algebraic techniques estimate the parameters of the close form corresponding to the global maximum, while adaptive techniques try to reach a local (hoped to be global) maximum by modifying the argument in a way that makes the objective increasing. In the BSS context, some contrast functions can be optimized via algebraic techniques (this is the case of criterion given in Eq. (1.36), whose optimization reduces to tensor diagonalization; the latter can be obtained via Jacobi techniques and, at each step, the angles are available in closed form [Cardoso and Souloumiac, 1993]), see e.g. [Cardoso and Comon, 1996] for a review. But for a wide class of contrasts, there exists no algebraic methods that would make possible the global maximization of the BSS criteria. Consequently, we need adaptive rules that will make \mathbf{B} converge to $\mathbf{B}^* \sim_u \mathbf{A}^{-1}$. The general form (gradient ascent) of these update rules, depending of the method of separation that has been chosen, is given below.

- Simultaneous separation of the K sources update, until convergence matrix \mathbf{B} subject to the constrain that $\mathbf{B}^{(t)} \in \mathcal{M}(K)$ holds at each step:

$$\mathbf{B}^{(t+1)} \leftarrow \mathbf{B}^{(t)} + \mu^{(t)} \Delta(\mathbf{B}^{(t)}). \quad (1.64)$$

- Sequential extraction of the K sources (deflation)

for i ranging from 1 to K , update until convergence the rows of \mathbf{B} subject to the constraint that $\mathbf{B}^{(t)} \in \mathcal{M}(K)$ holds at each step:

$$\mathbf{b}_i^{(t+1)} \leftarrow \mathbf{b}_i^{(t)} + \mu^{(t)} \Delta(\mathbf{b}_i^{(t)}). \quad (1.65)$$

- Partial separation of $P \leq K$ sources
update until convergence matrix \mathbf{B} such that $\mathbf{B}^{(t)} \in \mathcal{M}^{P \times K}$ holds at each step:

$$\mathbf{B}^{(t+1)} \leftarrow \mathbf{B}^{(t)} + \mu^{(t)} \Delta(\mathbf{B}^{(t)}) . \quad (1.66)$$

In the above update rules $\mu^{(t)}$ is a learning rate parameter. Obviously, one needs to find a suitable form for the pushforward term $\Delta(\cdot)$. Clearly, $\Delta(\cdot)$ must depend on the evolution of $\mathcal{C}(\cdot)$. If $\Delta(\cdot)$ has the form of the gradient of $\mathcal{C}(\cdot)$ with respect to the elements of \mathbf{B} , then the above rules are called *gradient ascent*.

- geodesic optimization on the $\mathcal{SO}(K)$ manifold (global or deflation)

In order to limit the computational load, one may deal with orthogonal BSS contrasts, and constrain \mathbf{B} to be always kept on $\mathcal{SO}(K)$; then, only $K(K-1)/2$ parameters have to be estimated. This is the so-called *geodesic search on the Stiefel manifold of special orthogonal matrices*. For instance, the above update rule for simultaneous separation must be modified such that for each t , $\mathbf{B}(t) \in \mathcal{SO}(K)$.

According to the group structure of $\mathcal{SO}(K)$, the aforementioned constraint always holds if, at each step t , the update rule is modified as

$$\mathbf{B}^{(t+1)} \leftarrow \mathbf{R}^{(t)} \mathbf{B}^{(t)} , \quad (1.67)$$

provided that $\mathbf{R}^{(t)} \in \mathcal{SO}(K)$ and $\mathbf{B}^{(0)} \in \mathcal{SO}(K)$.

Such a geodesic search can be done by using Jacobi rotations. Because of the group structure of $\mathcal{SO}(K)$ [Plumbley, 2004], for any pair of matrices \mathbf{B}, \mathbf{G} in $\mathcal{SO}(K)$ then $\mathbf{GB} \in \mathcal{SO}(K)$. Therefore a geodesic optimization can be obtained by factorizing \mathbf{B} as a product of rotation matrices, and we can choose $\mathbf{R}^{(t)} = \mathbf{G}_{ij}^{\alpha(t)}$ ($i < j$), where $\mathbf{G}_{ij}^{\alpha(t)}$ is a Givens matrix. A Givens matrix is a rotation matrix equal to the identity except entries $[\mathbf{G}_{ij}^{\alpha}]_{ii} = [\mathbf{G}_{ij}^{\alpha}]_{jj} = \cos \alpha$ and $[\mathbf{G}_{ij}^{\alpha}]_{ij} = -[\mathbf{G}_{ij}^{\alpha}]_{ji} = \sin \alpha$. At each step, the rotation angle $\alpha(t)$ is updated so that the criterion is increased.

1.7 BSS AND INFORMATION MEASURES

In the above subsection, various extraction schemes have been presented. As explained in the introduction of the section, the independence measure is a contrast for the simultaneous approach only. In order to obtain deflation and partial contrasts, we need another class of measures: the *measure of information*, a quantity that will be defined and explained below, is a possible candidate to derive contrast functions.

1.7.1 Information measure

The ICA approach to BSS tells us that, in order to solve the BSS problem in a simultaneous extraction scheme, one has to maximize an *independence measure*, as stated by Corollary 3 p. 20. Another viewpoint is the following. Assume that we can find a *complexity measure* of a signal, in the sense that the complexity measure of a linear combination of signals is larger than the smallest complexity measure of any of the individual signals, up to some normalization constraint on the mixture weights. Intuitively, if such a measure can be found, it is reasonable to think that minimizing the complexity measure of $\sum_{i=1}^K w_i S_i$, up to a normalization constraint of the form

$$\|\mathbf{w}\|_p \doteq \sqrt[p]{\sum_{i=1}^K w_i^p} = cst , \quad (1.68)$$

with respect to $\mathbf{w} = [w_1, \dots, w_K]$ would yield the source signal S_j with the lowest complexity measure. Yet another point of view is to consider the complexity measure as a sparsity measure of \mathbf{w} : the minimum complexity value is obtained when \mathbf{w} is the most sparse, i.e. has a single non-zero component, equal to the above constraint value fixing the p -norm of \mathbf{w} . Observe that throughout this thesis, $\|\mathbf{w}\|$ is most often used instead of $\|\mathbf{w}\|_2$, for short.

The complexity measure of a signal, as informally described above, can be thought of as a measure of the uncertainty of a process. It is a kind of *information measure* in the sense that intuitively, the more complex the signal, the more random the outcome; a large information is contained in the outcome since, for an observer, the outcome was not easily predictable. Consequently, a low information measure of a linear mixture of signals would thus mean that the mixture contains few “basis signals” with low uncertainty.

An information measure of a random variable could be, for example, the minimum number of bits needed to code the variable under constraint that the outcomes are one-to-one (i.e. univocally) decodable. The higher is the minimum number of bits required for the coding, the more complex is the underlying signal; it seems reasonable that coding $Y_i = \sum_{i=1}^K w_i S_i$ requires a larger number of bits than coding the “simplest” variable S_i if the S_i are independent random variables.

The information measure is the starting point of information theory, a field concerning the mathematical aspects of preserving, transforming and transmitting a message. Information measures are introduced in a very simple case in the next subsection, to yield then the *entropy* concept.

1.7.1.1 Discrete introductory example and Hartley's formula

This section is inspired from the following books: [Rényi, 1966, Cover and Thomas, 1991, MacKay, 2003].

Example 4 (Questions and Hartley's entropy) As an introduction to the information measure, assume that E_N is a discrete random variable with alphabet

$\mathcal{A}_E(N) = \{0, 1, \dots, N-1\}$ and that $\Pr(E_N = j) = 1/N$ for all $j \in \mathcal{A}_E(N)$, where $\Pr(\cdot)$ is a probability measure. We would like to know what is the information of an observation of E_N . Let us suppose $N = 8$. The information measure of E_N could be, for example, the minimum number of questions needed to find a given number, say n , in $\mathcal{A}_E(8)$. Actually, the best strategy is to ask the following questions [MacKay, 2003]:

- is $n \geq 4$?
- is $n \bmod 4 \geq 2$?
- is $n \bmod 2 = 1$?

Then the minimum number of questions needed to find the sought number is equal to 3. This corresponds to $\log_2 8$, which is Hartley's formula of the information amount of E_N ; $\log_2 N$ is the minimum number of bits required to code a number univocally in the set $\mathcal{A}_E(N)$ if the elements of $\mathcal{A}_E(N)$ are equiprobable.

According to Hartley, the information measure should satisfy ideally the following axioms:

- *Additivity*: the information measure of E_{NM} must be equal to the sum of the information measures of E_N and E_M . Indeed, the set $\mathcal{A}_E(NM)$ can be decomposed in N disjoint subsets $\mathcal{A}_E(M)^{(1)}, \dots, \mathcal{A}_E(M)^{(N)}$, each containing M elements. Finding a number $e \in \mathcal{A}_E(NM)$ can be managed by first finding the subset $\mathcal{A}_E(M)^{(j)}$ including e (requiring $\log_2 N$ questions, as given by Hartley's formula) and then finding the number in $\mathcal{A}_E(M)^{(j)}$ (requiring $\log_2 M$); the information measure of E_{NM} equals $\log_2(NM) = \log_2 N + \log_2 M$, and is thus additive.
- *Increasing with complexity*: The minimum number of bits required to code bi-univocally E_N increases with N ; the information measure of E_{N+1} is larger than or equal to that of E_N .
- *Normalization*: the information measure of E_2 is set to one, and this unit is named "bit", because it is the information measure contained in one bit.

It can be shown that $\log_2 N$ is the only functional satisfying the above axioms (p. 498 of [Rényi, 1966]).

The above information measure $\log_2 N$ equals $-\log_2 1/N$, i.e. minus the log of the probability that a uniform discrete random variable E_N with alphabet composed of N elements takes a specific value. When the random variable is not necessarily uniform and possibly continuous, the information measure is defined as follows.

Definition 20 (Information measure) *The information measure of a random variable X with probability mass function p_X is defined by the quantity*

$$\log 1/p_X. \quad (1.69)$$

It satisfies all the above axioms. The information unit of the amount of information is the “bit” (standing for BIrary digiT) if the base-2 is used; if the Natural logarithm is used instead, the unit is the “nat” (for NATural digiT).

1.7.1.2 Information and entropy

Assume that X_1, \dots, X_N are discrete i.i.d. random variables drawn from a pdf p_X with alphabet \mathcal{A}_X . We can define the average information measure $-1/N \sum_{i=1}^N \log p_{X_i}$. From the Asymptotic Equipartition Property (AEP) [Gray and Davisson, 2004, Gray, 1991, Cover and Thomas, 1991], this converges in probability to $-E[\log p_X]$ when $N \rightarrow \infty$. The quantity $-E[\log p_X]$ is called (discrete) Shannon’s entropy [Shannon, 1948]:

$$H[p_X] \doteq - \sum_{x \in \mathcal{A}_X} \Pr(X = x) \log \Pr(X = x) \quad (1.70)$$

$$= - \sum_{x \in \mathcal{A}_X} p_X(x) \log p_X(x) \quad (1.71)$$

$$= E[\log 1/p_X]. \quad (1.72)$$

If X is a random variable with pdf p_X , we note $H(X) = H[p_X]$.

Shannon’s entropy is thus the expected information measure of a random variable. It is a remarkable quantity satisfying, among others, the following:

- $H(\cdot) \geq 0$,
- $H(X_1, \dots, X_K) \leq \sum_{i=1}^K H(X_i)$ with equality if and only if X_i are independent,
- $H(X) \leq \log |\mathcal{A}_X|$ with equality if and only if p_X is the uniform pdf.

The following statistical meaning of H shows the key role played by this quantity:

Theorem 4 (Source Coding Theorem) *In average, if an experiment is repeated many times, we need more than $H[p_X]$ and only arbitrarily more than $H[p_X] + 1$ bits to code the results of an outcome of a random variable with pdf p_X .*

Shortly, the proof of this theorem relies on the fact that a one-to-one decodable binary code needs necessarily more than $H(X)$ bits (from Kraft inequality) and that there exists such a code requiring less than $H(X) + 1$ bits; this code assigns to each element $e_i \in \mathcal{A}_X$ a binary codeword of length $\lceil -\log_2 p_i \rceil$ where $p_i = \Pr(X = e_i)$. By doing so, the average length of a codeword is $-\sum_i p_i \lceil \log_2 p_i \rceil$. This is illustrated in a simple example from [Cover and Thomas, 1991].

Example 5 (Optimal coding) *Suppose we have a horse race with eight horses taking part (the alphabet is $\mathcal{A}_X = \{1, \dots, 8\}$), and assume that the probability of*

winning of these horses are given by the vector

$$\mathbf{p}_X = [1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64] .$$

We would like to find the optimal code for X , the winner of the race. The entropy of the race, which is also by the Source Coding Theorem the minimum number of bits to obtain a one-to-one decodable code for coding the outcome of X , is 2 bits. Clearly, $\log_2 \#\mathcal{A}_X = 3$ bits, which is the same number of bits as the one given by Hartley's formula, is a suboptimal coding scheme; attributing the same number of bits to each of the horses does not necessarily lead to minimize the code length of the winner of the race. By contrast, attributing a codeword length of $\lceil -\log_2 p_i \rceil$ bits to each horse would result in an optimal code. Such a code can be obtained by giving to each of the 8 horses, the codewords 0,01,001,0001,000000,000001,000010,000011 respectively. A shorter code is assigned to more probable outcomes. In average, the codeword lengths are equal to the entropy as in this example, $\lceil \log_2 p_i \rceil = \log_2 p_i$.

1.7.1.3 Extension to continuous random variables

The above section deals with discrete (alphabet) variables. However, the entropy concept also applies to continuous variables by replacing the sum symbols by integrals (in the sense of Riemann). This gives the so-called *differential entropy*:

$$\begin{aligned} h(X) = h[\mathbf{p}_X] &\doteq - \int_{\Omega(X)} \mathbf{p}_X(x) \log \mathbf{p}_X(x) dx \\ &= E[\log 1/\mathbf{p}_X] . \end{aligned} \quad (1.73)$$

The *differential entropy* is also called *Shannon's entropy* or abusively *entropy*, for short, when no confusion is possible. In spite of this apparent similarity between h and H , the latter has a rather different behavior. The differential entropy h is sensitive to the scale of the random variable; H was not since it only depends on the probability of the values of the random variable, which are not sensitive to the scale of the variable, and not on the possible values of the random variable itself. Similarly, while H is always positive, h may be negative, depending of the variance of the random variable. This directly results from the following property of the differential entropy:

Proposition 3 Let X be a continuous random vector with finite differential entropy and $Y = BX + \mu$ where B is a matrix and μ a vector. Then:

$$h(Y) = h(X) + \log |\det B| . \quad (1.74)$$

The entropy h is thus shift-invariant but scale sensitive. Observe that

$$h\left(\frac{X}{\sqrt{\text{Var}[X]}}\right) = h(X) - \frac{1}{2} \log \text{Var}[X] \quad (1.75)$$

is invariant under scaling.

This comes from the expression giving the density of a linear transformation of a random vector:

$$p_Y(Y) = \frac{1}{|\det \mathbf{B}|} p_X(\mathbf{B}^{-1}(Y - \mu)) . \quad (1.76)$$

Hence, provided that $|\det \mathbf{B}|$ is sufficiently close to zero (compared to the finite quantity $h(X)$), then $h(Y) < 0$. Similarly, provided that $|\det \mathbf{B}|$ is large enough, $h(Y)$ can be arbitrary large. However, under a power constraint on the random variable, it is possible to find the density with maximum (differential) entropy $h[\cdot]$.

Theorem 5 (Maximum entropy pdf) *Let $X \in \mathbb{R}^N$ be a zero-mean random vector with covariance matrix $\Sigma_X = \text{Cov}[X]$ and ϕ_X the multivariate Gaussian density of any zero-mean Normal vector of same dimension and covariance matrix as X . Then $h(X) \leq h[\phi_X]$ with equality if and only if $p_X = \phi_X$ almost everywhere.*

Proof: Let us note that

$$\phi_X(X) \doteq \frac{1}{(2\pi)^{N/2} \sqrt{\det \Sigma_X}} e^{\frac{-X^T \Sigma_X^{-1} X}{2}} . \quad (1.77)$$

Then:

$$0 \leq \text{KL}[p_X \| \phi_X] \quad (1.78)$$

$$= -h(X) - \int p_X \log \phi_X \quad (1.79)$$

$$\stackrel{(a)}{=} -h(X) - \int \phi_X \log \phi_X \quad (1.80)$$

$$= h[\phi_X] - h(X) . \quad (1.81)$$

The equality (a) results from the fact that $E_{p_X}[-X^T \Sigma_X^{-1} X] = E_{\phi_X}[-X^T \Sigma_X^{-1} X]$ implying $\int \phi_X \log \phi_X = \int p_X \log \phi_X$.

□

Simple algebraic manipulations show that $h[\phi_X] = \frac{1}{2} \log((2\pi e)^N \det \Sigma_X)$.

1.7.1.4 Information gain and Mutual information

It could be reasonable to understand the mutual information (MI)

$$\text{KL}(Y) = \text{KL} \left[p_Y \| \prod_{i=1}^K p_{Y_i} \right] \quad (1.82)$$

as the difference between the sum of information contained in each of the random variables Y_1, \dots, Y_K and the information contained in the joint set of these

random variables, that is, the information contained in the random vector $\mathbf{Y} = [\mathbf{Y}_1, \dots, \mathbf{Y}_K]^T$.

Then, for such an information measure, say $Q(\cdot)$, we could write

$$\text{KL}(\mathbf{Y}) = \sum_{i=1}^K Q(\mathbf{Y}_i) - Q(\mathbf{Y}) . \quad (1.83)$$

The functional Q can be found directly from the definition of the mutual information:

$$\begin{aligned} \text{KL} \left[p_{\mathbf{Y}} \parallel \prod_{i=1}^K p_{\mathbf{Y}_i} \right] &= \int p_{\mathbf{Y}_1, \dots, \mathbf{Y}_K}(y_1, \dots, y_K) \log p_{\mathbf{Y}_1, \dots, \mathbf{Y}_K}(y_1, \dots, y_K) dy_1 \dots dy_K \\ &\quad - \int p_{\mathbf{Y}_1, \dots, \mathbf{Y}_K}(y_1, \dots, y_K) \log \prod_{i=1}^K p_{\mathbf{Y}_i}(y_i) dy_1 \dots dy_K \end{aligned} \quad (1.84)$$

$$\stackrel{(a)}{=} E_{\mathbf{Y}}[\log p_{\mathbf{Y}}] - \sum_{i=1}^K \int_{y_i} p_{\mathbf{Y}_i}(y_i) \log p_{\mathbf{Y}_i}(y_i) dy_i \quad (1.85)$$

$$= \sum_{i=1}^K E_{\mathbf{Y}_i}[\log 1/p_{\mathbf{Y}_i}] - E_{\mathbf{Y}}[\log 1/p_{\mathbf{Y}}] . \quad (1.86)$$

Note that equality (a) comes from the marginalization on the joint density. From the above equation and Eq. (1.83), it results that we can choose $Q(\cdot) = h(\cdot)$ where h is the differential entropy, defined as in Eq. (1.73).

Another viewpoint is to consider $\text{KL}([\mathbf{X}, \mathbf{Y}])$ as the *gain of information* of one random variable resulting from the observation of the other. Let us define the conditional entropy of $\mathbf{Y} = [\mathbf{Y}_1, \dots, \mathbf{Y}_K]^T$ given \mathbf{Y}_k , $1 \leq k \leq K$ as

$$h(\mathbf{Y}|\mathbf{Y}_k) \doteq E_{\mathbf{Y}}[\log p_{\mathbf{Y}|\mathbf{Y}_k}] . \quad (1.87)$$

Then, by using the mathematical definition of the marginal, joint and conditional entropies, the mutual information between \mathbf{X}, \mathbf{Y} is

$$\begin{aligned} \text{KL}([\mathbf{X}, \mathbf{Y}]) &= h(\mathbf{X}) + h(\mathbf{Y}) - h(\mathbf{X}, \mathbf{Y}) \\ &= h(\mathbf{X}) - h(\mathbf{X}|\mathbf{Y}) \\ &= h(\mathbf{Y}) - h(\mathbf{Y}|\mathbf{X}) , \end{aligned} \quad (1.88)$$

where $h(\mathbf{X})$ (resp. $h(\mathbf{Y})$) represents the uncertainty on the outcome of \mathbf{X} (resp. \mathbf{Y}) and $h(\mathbf{X}|\mathbf{Y})$ (resp. $h(\mathbf{Y}|\mathbf{X})$) represents the uncertainty on the outcome of \mathbf{X} (resp. \mathbf{Y}) knowing the outcome of \mathbf{Y} (resp. \mathbf{X}). If the variables are independent, observing $\mathbf{Y} = y$ does not modify the uncertainty on \mathbf{X} so that the gain $\text{KL}([\mathbf{X}, \mathbf{Y}])$ is zero. The gain is positive otherwise. This generalizes to more than two random variables (chain rule for the entropy, resulting from the

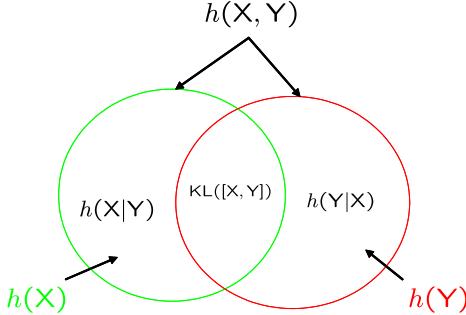


Figure 1.1. Venn diagram: relationships between entropies and mutual information.

definitions [Cover and Thomas, 1991]):

$$\begin{aligned}
 \text{KL}(\mathbf{Y}) &= \sum_{i=1}^K h(Y_i) - h(\mathbf{Y}) \\
 &= \sum_{i=1}^K h(Y_i) - (h(Y_1) + h(Y_2|Y_1) + h(Y_3|Y_1, Y_2) + \dots) \\
 &= \sum_{i=2}^K h(Y_i) - \sum_{i=2}^K h(Y_i|Y_1, \dots, Y_{i-1}) . \tag{1.89}
 \end{aligned}$$

This also shows the additivity nature of the entropy as an information measure: the information contained in the random vector \mathbf{Y} is the sum of the information contained in one of the K random variables, say Y_1 , plus the information brought by another one (say Y_2) given the first one, etc. In other words, the information brought by a new variable in the random vector equals the information of this variable knowing all the other ones already contained in \mathbf{Y} ; this information equals the information of the random variable if and only if it is independent from all the other components of the random vector. More explicitly, if Y_k is known, the remaining uncertainty on \mathbf{Y} reduces to $h(\mathbf{Y}|Y_k)$:

$$\begin{aligned}
 h(\mathbf{Y}) - h(\mathbf{Y}_k) &= -E_{\mathbf{Y}}[\log p_{\mathbf{Y}}] + E_{\mathbf{Y}}[\log p_{\mathbf{Y}_k}] \\
 &= E_{\mathbf{Y}}[\log \frac{p_{\mathbf{Y}_k}}{p_{\mathbf{Y}}}] \\
 &= -E_{\mathbf{Y}}[p_{\mathbf{Y}_1, \dots, \mathbf{Y}_{k-1}, \mathbf{Y}_{k+1}, \dots, \mathbf{Y}_K | \mathbf{Y}_k}(y_1, \dots, y_{k-1}, y_{k+1}, \dots, y_K | y_k)] \\
 &= h(\mathbf{Y} | \mathbf{Y}_k) \tag{1.90}
 \end{aligned}$$

Figure 1.1. gives a Venn diagram showing the relation between marginal and conditional entropies, joint entropy, and mutual information.

Remark 2 (Entropy, information and uncertainty) On the one hand $h(X)$ is the information measure of an outcome of X and, simultaneously, a

measure of the uncertainty of X ; how can that be possible? This is not, actually, a contradiction. Assume X is a Bernoulli random variable, taking the value 1 with probability p and 0 with probability $1 - p$, with $0 \leq p \leq 1$. Clearly, if $p = 0$ (resp. $p = 1$), $H(X) = 0$. The uncertainty contained in the variable is zero, which seems natural, and simultaneously, the outcome $X = 0$ (resp. 1) does not contain any information since there is no other possible value for the random variable. If p is close but different from 0 (resp. from 1), the uncertainty of X is low; we can predict, with a relatively high confidence, that $X = 0$ (resp. $X = 1$) will be observed: the information given by an outcome is thus quite useless, since one can guess it with a high confidence. If, on the contrary, $p = 1/2$, then it is very difficult to know in advance which value will be observed for X ; the information provided by an outcome of X is thus really useful since it is very difficult to predict. It is thus convenient to understand (intuitively) the “information measure” as the information needed to predict reliably the outcome of X ; the higher is the randomness of a system, the larger is, for an observer outside the studied physical system, the “lack of information about the state of the system” and thus the larger is the amount of information needed to guess the outcome of a given event; the entropy is precisely the averaged information. For more details about the meanings of entropy, we refer to [Brissaud, 2005, Arndt, 2004, Balatoni and Rényi, 1976].

Remark 3 The KL divergence (also called relative entropy) is a divergence measure between densities, but can also be seen as a relative information measure. From [Mourier, 1946], two densities differ more or less from each other according to how difficult it is to discriminate between them with the best test. Let H_i , $i \in \{1, 2\}$ be the hypothesis that x was drawn from the density p_i , then $\log \frac{p_1}{p_2}$ is the information in x for discrimination between H_1 and H_2 [Kullback and Leibler, 1951]. Hence, the mean information for discriminating between the H_i 's per observation from a subset $\mathcal{E} \subseteq \Omega_1$ is

$$\frac{\int_{x \in \mathcal{E}} p_1(x) \log \frac{p_1(x)}{p_2(x)} dx}{\int_{x \in \mathcal{E}} p_1(x) dx} \quad (1.91)$$

if $\int_{x \in \mathcal{E}} p_1(x) dx > 0$ and 0 otherwise. For $\mathcal{E} = \Omega_1$ we recover the KL:

$$\int_{x \in \Omega_1} p_1(x) \log \frac{p_1(x)}{p_2(x)} dx = \text{KL}[p_1 \| p_2] . \quad (1.92)$$

1.7.2 Entropy as a “complexity measure”

Let us now show that the entropy information measure can be seen as a complexity measure as defined in Section 1.7.1. The reasoning relies on the fundamental entropy power inequality (EPI).

Theorem 6 (Entropy Power Inequality (EPI)) Let S_1, S_2 be independent random variables with finite entropy h . Then

$$2^{2h(S_1+S_2)} \geq 2^{2h(S_1)} + 2^{2h(S_2)} , \quad (1.93)$$

with equality if and only if both S_i follow the Normal law.

This theorem first appeared in [Shannon, 1948], but it seems that the first formal proofs are due to [Stam, 1959] and [Blachman, 1965]. It is also interesting to have a look at [Verdu and Guo, 2006] in which a very simple proof of this theorem, exploiting the relationships between mutual information and minimum mean-square error in Gaussian channels, has been recently proposed.

It comes immediately that the following corollary holds:

Corollary 4 Let S_1, \dots, S_K be independent random variables with finite entropy and $K > 1$. Then

$$2^{2h(\sum_{i=1}^K S_i)} \geq \sum_{i=1}^K 2^{2h(S_i)}, \quad (1.94)$$

with equality if and only if all the S_i follow the Normal law.

Proof: The proof of this corollary consists in first observing that $2^{2h(\sum_{i=1}^K S_i)} \geq 2^{2h(S_1)} + 2^{2h(\sum_{i=2}^K S_i)}$ since S_1 is independent from $\sum_{i=2}^K S_i$ (by the converse form of the Darmois-Skitovitch Theorem, Theorem 2 p. 12). The equality is attained only if S_1 and $\sum_{i=2}^K S_i$ are Normal random variables. But from the EPI, $2^{2h(\sum_{i=2}^K S_i)} \geq 2^{2h(S_2)} + 2^{2h(\sum_{i=3}^K S_i)}$ for the same reason as above with equality if and only if both S_2 and $\sum_{i=3}^K S_i$ are Gaussian random variables, leading to $2^{2h(\sum_{i=1}^K S_i)} \geq 2^{2h(S_1)} + 2^{2h(S_2)} + 2^{2h(\sum_{i=3}^K S_i)}$ in which the equality holds true if and only if S_1, S_2 and $\sum_{i=3}^K S_i$ are Gaussian. One concludes the proof by iterating this result (by recurrence).

□

This theorem is the keystone for proving the intuitive nature of the *complexity measure* of the entropy, stated in the following lemma.

Let us define $\mathcal{S}(K)$ to be the set of K -entries unit-norm vectors:

$$\mathcal{S}(K) \doteq \{\mathbf{w} \in \mathbb{R}^K : \|\mathbf{w}\| = 1\}. \quad (1.95)$$

Lemma 5 Consider the K -dimensional vector $\mathbf{w} \in \mathcal{S}(K)$ and $\mathbf{S} = [S_1, \dots, S_K]^T$, where at most one of the sources is Gaussian. Then, the global minimum of $h(\mathbf{w}\mathbf{S})$ is reached when and only when $\mathbf{w} = \pm \mathbf{e}_k$, $k \in \operatorname{argmin}_{i \in \{1, \dots, K\}} h(S_i)$.

Proof: The proof consists of two parts. First, assume that $\|\mathbf{w}\|_\infty = \|\mathbf{w}\|$, that is, $\mathbf{w}\mathbf{S}$ is equal to one source (up to its sign), say S_j . Then, clearly, $h(\mathbf{w}\mathbf{S}) = h(S_j)$; this quantity is minimum if j corresponds to the index of one of the sources with minimum entropy.

Suppose now that at least two entries of \mathbf{w} are non-zero, and that $I(\mathbf{w})$ is the set of the indexes of these non-zero elements:

$$I(\mathbf{w}) \doteq \{i \in \{1, \dots, K\} : w_i \neq 0\}, \quad (1.96)$$

with $\#[I(\mathbf{w})] \geq 2$. Then, noting that at most one source is Gaussian, the strict inequality holds true in Corollary 4 and we get

$$\begin{aligned} 2^{2h(\mathbf{w}\mathcal{S})} &= 2^{2h(\sum_{i \in I(\mathbf{w})} w_i \mathcal{S}_i)} > \sum_{i \in I(\mathbf{w})} 2^{2h(w_i \mathcal{S}_i)} \\ &\stackrel{(a)}{=} \sum_{i \in I(\mathbf{w})} 2^{2(h(\mathcal{S}_i) + \log |w_i|)} \\ &= \sum_{i \in I(\mathbf{w})} 2^{2h(\mathcal{S}_i)} 2^{\log w_i^2} \\ &= \sum_{i \in I(\mathbf{w})} w_i^2 2^{2h(\mathcal{S}_i)}, \end{aligned} \quad (1.97)$$

where (a) results from Property 1.74 (p. 33). Clearly, since the logarithm is a strictly increasing function, the last expression reduces to

$$h(\mathbf{w}\mathcal{S}) > \frac{1}{2} \log_2 \left(\sum_{i \in I(\mathbf{w})} w_i^2 2^{2h(\mathcal{S}_i)} \right). \quad (1.98)$$

But since $\|\mathbf{w}\| = 1$, we have

$$w_j^2 = 1 - \sum_{i \in I(\mathbf{w}) \setminus \{j\}} w_i^2 \quad (1.99)$$

for all $j \in I(\mathbf{w})$, and in particular for $j = k'$ where

$$k' \in \operatorname{argmin}_{i \in I(\mathbf{w})} h(\mathcal{S}_i). \quad (1.100)$$

Then, it comes by definition of k'

$$\begin{aligned} \sum_{i \in I(\mathbf{w})} w_i^2 2^{2h(\mathcal{S}_i)} &= 2^{2h(\mathcal{S}_{k'})} + \underbrace{\sum_{i \in I(\mathbf{w}) \setminus \{k'\}} w_i^2 (2^{2h(\mathcal{S}_i)} - 2^{2h(\mathcal{S}_{k'})})}_{\geq 0} \end{aligned} \quad (1.101)$$

and finally

$$\begin{aligned} h(\mathbf{w}\mathcal{S}) &> \frac{1}{2} \log_2 \left(2^{2h(\mathcal{S}_{k'})} \right) \\ &= h(\mathcal{S}_{k'}) \\ &\geq h(\mathcal{S}_k) \end{aligned} \quad (1.102)$$

since $I(\mathbf{w}) \subseteq \{1, \dots, K\}$.

□

Clearly, Lemma 5 is the starting point for looking at BSS criteria based on information measures. They could constitute a wide class of BSS contrasts. We shall focus on information measures called *r-th order information measure* or *Rényi's information measure*. The definition of this class of information measures is the purpose of the next subsection.

1.7.3 Generalized information measures

Rényi's entropy is a generalization of Shannon's in the sense that both share the same key properties of information measures [Rényi, 1976a]. It is defined as

$$H_r[p_X] \doteq \frac{1}{1-r} \log \left(\sum_{x \in \mathcal{A}_X} p_X^r(x) \right) , \quad (1.103)$$

where $r \geq 0$ (the non-negativity of Rényi's exponent is always assumed throughout this work even if not explicitly mentioned). As for Shannon's entropy, we note $H_r(X) = H_r[p_X]$ if X follows the density p_X . To see where does this extended form of information measure come from, observe that in the general theory of means [Aczel, 1948, Hardy et al., 1934], the mean of the real numbers x_1, \dots, x_n with respective weights w_1, \dots, w_n ($w_i > 0$, and $\sum_{i=1}^n w_i = 1$) is an expression of the form

$$\vartheta^{-1} \left(\sum_{i=1}^n w_i \vartheta(x_i) \right) , \quad (1.104)$$

where the usual definition of mean of the x_i is obtained for ϑ being any linear function and $w_i = 1/N$. Hence, the general average of information measure, noting $p_i \doteq \Pr(X = x_i)$ is

$$\vartheta^{-1} \left(\sum_{i=1}^n p_i \vartheta(\log 1/p_i) \right) . \quad (1.105)$$

But in order to preserve the additivity property of the average information measure axiom, ϑ cannot be arbitrary. It can obviously be linear (it corresponds then to Shannon's entropy H defined in Eq. (1.72)), but it can also be an exponential functional [Rényi, 1976a]. The quantity H_r is obtained by taking $\vartheta(x) = 2^{(1-r)x}$ or $\vartheta(x) = e^{(1-r)x}$ depending on the log being the base-2 or natural logarithm (i.e. if the entropy is expressed in bits or in nats).

Just as for Shannon, we can extend the discrete Rényi entropy to continuous densities:

$$h_r[p_X] \doteq \frac{1}{1-r} \log \int_{\Omega(X)} p_X^r(x) dx , \quad (1.106)$$

where $r \geq 0$ and $\Omega(X)$ is the support (set) of the random variable X . As usual, we note $h_r(X) = h_r[p_X]$ if X has the density p_X . Rényi's entropy satisfies the following [Cover and Thomas, 1991]:

- $\lim_{r \rightarrow 1} h_r(X) = h_1(X) = h(X)$;
- $\lim_{r \rightarrow 0} h_r(X) = h_0(X) = \log \mu[\Omega(X)]$ where $\mu[\cdot]$ denotes the Lebesgue measure;
- Rényi's entropy is continuous and decreasing in r [Lutwak et al., 2005, Ben-Bassat and Raviv, 1978];
- As for Shannon's entropy, if α, β are two scalars and $\mathbf{M}_1, \mathbf{M}_2$ two $K \times K$ matrices:

$$\begin{cases} h_r(\alpha X + \beta) = h_r(X) + \log |\alpha| , \\ h_r(\mathbf{M}_1 X + \mathbf{M}_2) = h_r(X) + \log |\det \mathbf{M}_1| . \end{cases}$$

Proof: It is easily checked that Rényi's entropy is not sensitive to translation. Regarding the scaling, we have from Eq. (1.76)

$$\begin{aligned} h_r(\mathbf{M}_1 X) &= \frac{1}{1-r} \log \left(\int_{\Omega(\mathbf{M}_1 X)} p_{\mathbf{M}_1 X}^r(\mathbf{M}_1 X) d(\mathbf{M}_1 X) \right) \\ &= \frac{1}{1-r} \log \left(\int_{\Omega(X)} \frac{1}{|\det \mathbf{M}_1|^r} p_X^r(X) \det \mathbf{M}_1 dX \right) \\ &\stackrel{(a)}{=} \frac{1}{1-r} \log \left(|\det \mathbf{M}_1|^{1-r} \int_{\Omega(X)} p_X^r(X) dX \right) \\ &= \log |\det \mathbf{M}_1| + h_r(X) . \end{aligned} \quad (1.107)$$

Note that equality (a) holds true even when $\det \mathbf{M}_1 < 0$ since in this case, $\det \mathbf{M}_1$ is multiplied by -1 when the ad-hoc bounds of the integral have been suitably swapped.

□

As a matter of fact, just as for Shannon's entropy (see Eq. (1.75)):

$$h_r(X / \sqrt{\text{Var}[X]}) = h_r(X) - \frac{1}{2} \log \text{Var}[X] \quad (1.108)$$

is a scale-invariant function of the random variable X .

In the sequel, Rényi's entropy is either denoted $h_r(X)$ or $h_{r,\Omega}(X)$, Ω being the support set of the random variable argument X (instead of $\Omega(X)$ for short, when no confusion is possible). It is possible to define an extended form of Rényi's entropy, called *Extended Rényi's Entropy* (ERE), noted $h_{r,\bar{\Omega}}(X)$, which is defined as Rényi's entropy except that the integration domain in Eq. (1.106)

is the convex hull $\bar{\Omega}$ of the support Ω , that is the smallest convex set including Ω :

$$h_{r,\bar{\Omega}}(X) \doteq \frac{1}{1-r} \log \int_{\bar{\Omega}(X)} p_X^r(x) dx . \quad (1.109)$$

However, the pdf is undefined out of $\Omega(X)$. In order to deal with densities on the whole real line, we set $p_X(x) = 0$ for all $x \in \mathbb{R} \setminus \Omega(X)$. Clearly, the following corollary holds true.

Corollary 5 *The extended Rényi entropy satisfies the following:*

$$\begin{aligned} h_{r,\bar{\Omega}}(X) &= h_{r,\Omega}(X) = h_r(X) \text{ if } r > 0 \\ h_{0,\bar{\Omega}}(X) &\geq h_{0,\Omega}(X) = h_0(X) \text{ with equality if and only if } \mu[\bar{\Omega}(X)] = \mu[\Omega(X)]. \end{aligned}$$

Proof The proof is straightforward. If $r > 0$, then, if the base-2 for the logarithm is used:

$$\begin{aligned} h_{r,\bar{\Omega}}(X) &= \frac{1}{1-r} \log \left\{ \int_{\Omega(X)} p_X^r(x) dx + \int_{\bar{\Omega}(X) \setminus \Omega(X)} p_X^r(x) dx \right\} \\ &= \frac{1}{1-r} \log \left\{ 2^{(1-r)h_r(X)} + \int_{\bar{\Omega}(X) \setminus \Omega(X)} 0^r(x) dx \right\} \\ &= h_{r,\Omega}(X) = h_r(X) . \end{aligned} \quad (1.110)$$

On the other hand, $h_0 = \log \mu[\Omega(X)]$ and

$$\begin{aligned} h_{0,\bar{\Omega}}(X) &= \log \left\{ \mu[\Omega(X)] + \int_{\bar{\Omega}(X) \setminus \Omega(X)} dx \right\} \\ &= \log \mu[\bar{\Omega}(X)] , \end{aligned} \quad (1.111)$$

which is greater than $h_0(X) = h_{0,\Omega}(X)$ with equality if and only if $\mu[\bar{\Omega}(X)] = \mu[\Omega(X)]$ as the log function is monotonic.

□

For more information on Rényi's entropy, we refer to the monograph of [Aczel and Daroczy, 1975].

1.8 ISSUES AND OBJECTIVES OF THE THESIS

In this chapter, the BSS task has been mathematically written and its solutions were formulated in terms of non-mixing matrices. After having briefly recalled the relationships between independence (and thus ICA) and BSS, it was explained that other contrast functions are needed, especially regarding the partial and deflation procedures. Information measures derived from Rényi's entropies

seem to be interesting candidates, as shown for the specific Shannon entropy case. The particular Shannon entropy has close relationships to mutual information (as shown by Eq. (1.88)) and non-Gaussianity approaches, and have been already suggested for BSS. Sometimes however, only informal arguments are used for justifying the use of entropy as a BSS contrast function for e.g. deflation (see Section 2.2.2).

We are now able to introduce the original contributions of the thesis. Based on Lemma 5, it has been explained that the opposite of Shannon's information measure is a good candidate for a contrast function (to be maximized); but what about the more general class of Rényi's entropies? This is a natural question as it shares the same information measure properties than Shannon's one. In other words, we shall analyze if, generally speaking, the opposite of information measures such as $h_r(X)$ are contrast functions for simultaneous, deflation or partial separation. Chapter 2 proposes a unifying investigation of the contrast properties of $h_r(X)$, and formal proofs are provided when possible. Even if some results exist regarding the contrast properties of Shannon's entropy, the particular cases $r = 1$ and $r = 0$ of Rényi's entropy will be considered before trying to generalize the results to h_r . The reason for doing that is threefold:

- for being self-complete regarding entropic contrast functions,
- for giving alternative (and simpler) proofs of the contrast properties,
- for developing proof strategies that are extendable to the general case.

In summary, the next chapter will tell us if the *global maxima* of the criteria based on h_r i) if $r = 1$, ii) $r = 0$ or iii) in the general case $r > 0$ yield the sought sources, whatever the extraction scheme. An additional study is further provided for the deflation and partial schemes: the analysis of the possible *local maxima* of the criteria when the demixing matrix satisfies $\mathbf{B} \sim_u \mathbf{A}^{-1}$, that is when $\mathbf{BA} \in \mathcal{W}^{P \times K}$; they are called *non-mixing maxima*. For instance, does the contrast function has a local maximum point when any (subset of) the K sources is extracted?

Chapter 3 will tackle another problem related to the BSS contrast functions. It was explained in Section 1.6 that when maximizing some particular criteria (such as the ones based on information-theoretic criteria), non-algebraic (i.e. iterative) optimization techniques similar to those given in the last section have to be used. The problem is that these rules converge to a *local maximum* (if it exists) of the function. But according to its definition, the *global maximum* of a BSS contrast must be reached in order to recover $\mathbf{B} \sim \mathbf{A}^{-1}$. Therefore, the maximization algorithm may be stuck in a so-called *mixing maximum*, that is a local maximum that does not correspond to an acceptable solution of the BSS problem; the obtained matrix \mathbf{B} is not in the set of $P \times K$ non-mixing matrices $\mathcal{W}^{P \times K}$. Then, it is important to study the possible existence of the local maxima to know if we can "blindly trust" the solution obtained by the iterative maximization algorithm, that is if we have 100% confidence in the fact

that the algorithm, that will converge surely to a local maximum, will converge to a *non-mixing* point. The only way to be sure of the non-mixing specificity of the attained maximum points is to use a criterion that has no mixing local maximum. The BSS contrast functions that benefit from this nice behavior will be called in the following “discriminant” BSS contrasts.

The last analysis will reveal a major advantage of $h_{0,\bar{\Omega}}(Y_i)$ regarding the ERE with other values of r ; the latter criterion reduces actually to the log-range of the output Y_i . Because of its theoretical and practical advantages, the use of this criterion, as well as its estimation and extension to challenging BSS applications will finally be addressed in Chapter 4.

CHAPTER 2

CONTRAST PROPERTY OF ENTROPIC CRITERIA

ANALYSIS OF THE NON-MIXING MAXIMA

Abstract. In this chapter, we are interested in analyzing the contrast property of generalized information measures that have been proposed in the literature for solving the BSS problem. More specifically, we focus on the (possibly extended form) of Rényi's entropies, noted h_r (Section 1.7.3, Eq. (1.106)). The analysis in this chapter focuses on the *non-mixing maxima* of these criteria as a function of the demixing matrix elements. Two kinds of non-mixing maxima are analyzed:

- First, the *global* maximum points of the criteria are analyzed. These points are related to the contrast function property in the sense that they should correspond to transfer matrices $\mathbf{W} = \mathbf{B}\mathbf{A}$ that are non-mixing ($\mathbf{W} \sim_u \mathbf{I}_K$).

Some of them are already known but we remind them here, as well as some tools for proving the contrast property easily, when possible. In some cases, counter-examples are used to show that a given criterion is not a contrast function.

- Second, we focus on the less known *local* non-mixing maximum points of the criteria.

Yet another kind of maximum points exists: the *mixing maximum points* (corresponding to transfer matrix $\mathbf{W} \not\sim_u \mathbf{I}_K$). They shall be investigated separately in Chapter 3.

Contribution. Author's contribution is divided in two classes, for clarity. First, the results about the contrast properties of entropic criteria are summarized. Next, the mathematical tools that have been developed in order to perform the above analysis are listed.

- Results about the local and global optima of entropic criteria
 - **Shannon's entropy** was proved to lead to a contrast function for simultaneous separation (see [Comon, 1994]). It was then used in a deflation scheme, but we were not able to find a pioneering reference. We note that Hyvärinen proved in 1998 that this method sounds for specific approximations of Shannon's entropy [Hyvärinen, 1997]. This is proved here for the exact Shannon entropy based on the EPI. Similarly, Shannon's entropy is proved to have a local minimum point under a fixed variance constraint when this output is proportional to a non-Gaussian source, based on a Taylor expansion. Pham showed that the partial separation criterion reaches a stationary point when a subset of sources is extracted [Pham, 2006a].
 - The **range-based criterion** was proposed to be a contrast for simultaneous separation in [Pham, 2000], and then for deflation under prewhitening (orthogonal deflation contrast) [Cruces and Duran, 2004]. It is shown here that the range yields contrast functions for the three deflation schemes, even without prewhitening. Furthermore, under a fixed variance constraint on the outputs, it has a local minimum point when the output is proportional to a source. In partial separation scheme too, the related contrast reaches a local maximum point when the rectangular demixing matrix is subPD-equivalent to the inverse of the mixing matrix.
 - In a more general way and in a mimetic manner compared to Shannon's entropy, **Rényi's entropies** were conjectured to be contrast functions [Erdogmus et al., 2002a]. It is proved here that they are not, generally speaking, contrast function if Rényi's exponent is not set to zero or one, neither for deflation separation, nor for simultaneous separation. By contrast, we show that Rényi's entropy admits, under a fixed variance constraint, a stationary point when the output is proportional to a source; but the kind of this stationary point (minimum/maximum) depends on the value of the Rényi's exponent and on the source densities as well.
- Tools and other results
 - A Taylor expansion of Rényi's entropy is proposed;
 - Rényi's entropy is proved to not be a superadditive functional: some counter-examples of source density exist for every value of Rényi's exponent (other than 0 and 1) within the exponential family, even in the simple case involving two sources sharing a same density;

- An extended form of the Brunn-Minkowski inequality was given (not in the sense of the dimensionality, but in the sense of the “iff” statement for non-convex sets).

In this chapter, all the proofs are original. Some of them result from joint work with D.-T. Pham.

Part of the material presented in this chapter was or will be published in the following papers (see Appendix B, p. 279): JP1, JA1 (results about Shannon’s entropy) JA2, JA3, ICB6, ICP10(results about the support and the range) JS1, ICTBS1 (results about Rényi entropies).

Organization of the chapter. The chapter is organized as follows. After having reminded useful results for building BSS contrast functions, we study two specific cases of extended Rényi’s entropy (ERE): Shannon’s entropy ($r = 1$) and the Lebesgue log-measure of the support convex hull, also known as the “range” ($r = 0$); they are addressed in Section 2.2 and Section 2.3 respectively. Then, in Section 2.4, the extended Rényi entropy (ERE) is analyzed in its generalized form, without a priori fixing the value of $r > 0$, $r \neq 1$. The proofs are relegated to an appendix at the end of the Chapter for clarity (Section 2.6).

2.1 SOME TOOLS FOR BUILDING CONTRAST FUNCTIONS

The first method that comes in mind for showing that a criterion is a BSS contrast function for simultaneous, deflation or partial separation is to look at the contrast function definition, and to prove that the global maximum of the criterion corresponds to non-mixing transfer matrices. From Figure 2.1., this is equivalent to check if a given functional f is in the set \mathbb{F}_C defined as the set of functions matching the contrast function property. However, this might be quite heavy in some cases. An alternative approach consists in verifying other conditions implying the contrast property. For instance, a given functional is a contrast if *sufficient conditions* that guarantee that the contrast property holds are met. This is equivalent to checking if f is in a set $\tilde{\mathbb{F}}_C$ satisfying $\tilde{\mathbb{F}}_C \subset \mathbb{F}_C$. The advantage of the second approach compared to the first one is that the latter condition might be easier to check even if, in counter part, the conditions are unnecessarily strong. This is illustrated in Figure 2.1.

The next subsection gives two results ensuring that specific functionals are BSS contrast functions without analyzing the global maximum point of the functionals.

2.1.1 From orthogonal deflation to orthogonal partial separation

Under some conditions, one can guarantee that a sum of deflation (D-BSS) contrasts yields partial (P-BSS) and/or simultaneous (S-BSS) contrasts. This was stated and proved in [Cruces et al., 2004].

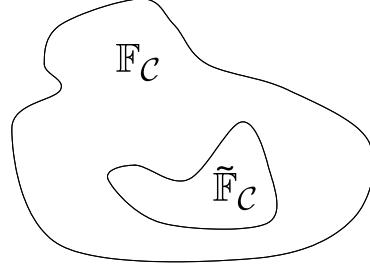


Figure 2.1. If $f \in \tilde{\mathbb{F}}_C$ then $f \in \mathbb{F}_C$ and the functional is a contrast function as sufficient conditions are met.

Theorem 7 (Cruces, Cichocki and Amari) Let us give K constants $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_K$ and a deflation contrast $\mathcal{C}(\cdot)$ (in the sense of Def. 13, p. 21) satisfying the following properties:

- $\mathcal{C}(\mathbf{b}) \geq 0$ with equality if and only if $\mathbf{b}\mathbf{X}$ is Gaussian;
- $\mathcal{C}(\cdot)$ satisfies a weak form of strict convexity: if $\mathbf{Y}_1 = \sum_{i=1}^K W_{1i} \mathbf{S}_i$ and $\sum_{i=1}^K W_{1i}^2 = 1$, then

$$\mathcal{C}(\mathbf{b}_1) \leq \sum_{i=1}^K W_{1i}^2 \mathcal{C}(\mathbf{e}_i \mathbf{A}^{-1}) \quad (2.1)$$

where, for $\mathcal{C}(\mathbf{b}_1) > 0$, the equality holds true if and only if $\mathbf{b}_1 \propto \mathbf{e}_i \mathbf{A}^{-1}$.

Assume further that the sources are ordered with respect to this contrast as

$$\mathcal{C}(\mathbf{e}_1 \mathbf{A}^{-1}) \geq \dots \geq \mathcal{C}(\mathbf{e}_P \mathbf{A}^{-1}) > \mathcal{C}(\mathbf{e}_{P+1} \mathbf{A}^{-1}) \geq \dots \geq \mathcal{C}(\mathbf{e}_K \mathbf{A}^{-1}) \quad (2.2)$$

Then, if $\mathcal{C}(\mathbf{e}_P \mathbf{A}^{-1}) > 0$, the objective function

$$\mathcal{C}(\mathbf{B}) = \sum_{i=1}^P \alpha_i \mathcal{C}(\mathbf{b}_i) \text{ subject to } \text{Cov}[\mathbf{Y}] = \mathbf{I}_P \quad (2.3)$$

is a P -BSS contrast function whose global maxima correspond to the extraction of the first P sources from the mixture. If, additionally, $\mathcal{C}(\mathbf{e}_1 \mathbf{A}^{-1}) > \mathcal{C}(\mathbf{e}_2 \mathbf{A}^{-1}) > \dots > \mathcal{C}(\mathbf{e}_P \mathbf{A}^{-1})$ and $\alpha_1 > \alpha_2 > \dots > \alpha_P$ then the global maximum is unique and corresponds to the ordered extraction of the first P sources of the mixture, i.e. the global maximum yields $\mathbf{Y} = [\mathbf{S}_1, \dots, \mathbf{S}_P]^T$ if \mathcal{A}_7 (i.e. $\text{Cov}[\mathbf{S}] = \mathbf{I}_P$) holds.

Several comments can be formulated about this theorem.

Remark 4 Note that it is further assumed in this theorem that the mixing matrix \mathbf{A} is orthogonal; this is equivalent to suppose that the mixtures have been whitened

by a whitening matrix \mathbf{V} and that \mathbf{A} has been replaced by \mathbf{VA} . In other words, the obtained partial contrast is actually an orthogonal partial contrast since:

$$\text{Cov}[\mathbf{Y}] = \mathbf{I}_P \quad (2.4)$$

$$= E[\mathbf{BAS}(\mathbf{ABS})^T] \quad (2.5)$$

$$= \mathbf{B} \underbrace{\text{Cov}[\mathbf{X}]}_{\mathbf{I}_K} \mathbf{B}^T \quad (2.6)$$

$$= \mathbf{BB}^T. \quad (2.7)$$

Clearly, the obtained criterion is an orthogonal S-BSS contrast if $P = K$.

Remark 5 Note that if the α_i take different values, the obtained “contrast” becomes sensitive to permutation. Observe further that the permutation problem is not solved, it is only constrained to be no more arbitrary: the sources S_i are ordered with respect to their deflation contrast value, not necessarily up to their initial order in S (if this notion makes sense, which is not clear at all as it is a simple mathematical notation). Finally, regarding the scale indeterminacy, we remind that it is only avoided because the sources are supposed to be known (by A_7).

Remark 6 We should probably point out the fact that the equality “ $\mathbf{Y} = [S_1, \dots, S_P]^T$ ” is too restrictive. If the contrast is sign-invariant, if a global maximum exists at $\mathbf{Y}^* = [S_1, \dots, S_P]^T$, then a local maximum should also exist at e.g. $-\mathbf{Y}^*$, provided that such an output can be obtained in spite of the $\mathbf{BB}^T = \mathbf{I}_P$ constraint. Clearly, this is the case. An identity matrix of order P in which a number of rows have been sign-inverted, noted \mathbf{M} , would satisfies $\mathbf{MB}(\mathbf{MB})^T = \mathbf{I}_P$ since $\mathbf{M} \in \mathcal{O}(P)$. Clearly, each component of \mathbf{BX} equals a component of \mathbf{MBX} , but possibly only up to its sign.

Remark 7 It results from a quick look at the proof of the above theorem that $\mathcal{C}(\mathbf{e}_P \mathbf{A}^{-1}) > \mathcal{C}(\mathbf{e}_{P+1} \mathbf{A}^{-1})$, a condition which indicates that P cannot be arbitrary chosen in $\{1, \dots, K\}$, is not necessary for the obtained global criterion to be a contrast. If this inequality is not strict, the global maximum is attained if the P sources with the larger value of the criterion are extracted. The estimated sources are ordered with respect to their deflation contrast value; any pair of sources sharing a same value of the deflation contrast can be permuted without affecting the value of the global contrast.

2.1.2 Huber's superadditivity concept: a simple tool for building simultaneous and partial contrast functions

An additional interesting result, due to Pham in 2001, exists regarding the contrast function property of a criterion:

Theorem 8 (Pham [2001a, 2006a]) Suppose that Q is a class II superadditive functional in the sense of Huber [Huber, 1985], i.e. that for any pair of independent random variables X, Y and two scalar numbers α, β :

$$\begin{cases} Q(\alpha X + \beta) = |\alpha|Q(X) , \\ Q^2(X + Y) \geq Q^2(X) + Q^2(Y) \end{cases} \quad \begin{array}{l} (\text{Huber 1}) \\ (\text{Huber 2}) \end{array}$$

and the strict equality holds in the last expression if and only if X and Y are Gaussian. Then, any criterion of the form

$$f^\square(\mathbf{B}) \doteq \log |\det \mathbf{B}| - \sum_{i=1}^K \log Q(\mathbf{b}_i \mathbf{X}) , \quad (2.8)$$

\mathbf{b}_i being the i -th row of \mathbf{B} , is a contrast function over the set $\mathcal{M}(K)$ of full-row rank $K \times K$ matrices for simultaneous separation. In other words, it reaches its global maximum point if and only if $\mathbf{BA} \in \mathcal{W}(K)$ or, equivalently, if and only if $\mathbf{B} \sim \mathbf{A}^{-1}$. Similarly, under the same condition on $Q(\cdot)$, any functional $f(\mathbf{B})$ of the form

$$f(\mathbf{B}) \doteq \frac{1}{2} \log |\det(\mathbf{B}\Sigma_X \mathbf{B}^T)| - \sum_{i=1}^P \log Q(\mathbf{b}_i \mathbf{X}) , \quad (2.9)$$

where $P \leq K$, $\Sigma_X \doteq \text{Cov}[\mathbf{X}]$ is the covariance matrix of $\mathbf{X} = \mathbf{AS}$, is a partial contrast function over the set $\mathcal{M}^{P \times K}$ of full-row rank $P \times K$ matrices; it reaches a global maximum point only if $\mathbf{BA} \in \mathcal{W}^{P \times K}$ or equivalently, only if $\mathbf{B} \sim_u \mathbf{A}^{-1}$.

Remark 8 Note that in the above theorem, it is implicitly required that $Q(\cdot)$ must be strictly positive in order that $\log Q(\cdot)$ exists, and $\max(Q(X), Q(Y)) < \infty$.

Remark 9 Observe that in Eq. (2.9), one may freely replace $|\det(\mathbf{B}\Sigma_X \mathbf{B}^T)|$ by $\det(\mathbf{WW}^T)$ with $\mathbf{W} \doteq \mathbf{BA}$ since

$$\mathbf{B}\Sigma_X \mathbf{B}^T = \mathbf{BA}\Sigma_S(\mathbf{BA})^T \stackrel{\mathcal{A}_7}{=} \mathbf{WW}^T , \quad (2.10)$$

and because \mathbf{WW}^T is positive definite and the determinant of a positive definite matrix is always positive.

It is easy to further characterize the set of the global maximum points of $f(\mathbf{B})$ (the first “if” in the “if and only if” expression is missing in the second claim of the theorem): from \mathcal{A}_7 (i.e. $\Sigma_S = \mathbf{I}$) and if the sources are ordered according to $Q(\cdot)$ as

$$Q(S_1) \leq Q(S_2) \leq \dots \leq Q(S_K) , \quad (2.11)$$

for the sake of simplicity, one gets the following corollary (the proof is given in the Appendix of the Chapter, Section 2.6.1, p. 84).

Corollary 6 (Characterization of global maximizer set of f) Let us define $P^m \doteq \min\{i \in \{1, \dots, P\} : Q(\mathbf{S}_i) = Q(\mathbf{S}_P)\} - 1$, and $P^M \doteq \max\{i \in \{P, \dots, K\} : Q(\mathbf{S}_i) = Q(\mathbf{S}_P)\}$. The global maximum points of f over the set $\mathcal{M}^{P \times K}$ are the matrices \mathbf{B} such that $\mathbf{B}\mathbf{A} \in \mathcal{W}_P^{P \times K}$, where $\mathcal{W}_P^{P \times K} \subset \mathcal{W}^{P \times K}$ is the set of matrices with exactly one non-zero element per row, at most one non-zero element per column and with P^m rows having a single non-zero element with column index in $\{1, \dots, P^m\}$. The remaining rows have a single non-zero element with column index in $\{P^m + 1, \dots, P^M\}$.

Example 6 Assume $K = 5$, $P = 2$ and $Q(\mathbf{S}_1) < Q(\mathbf{S}_2) = Q(\mathbf{S}_3) < Q(\mathbf{S}_4) \leq Q(\mathbf{S}_5)$. Then, $P_m = 1$, $P_M = 3$ and if we define

$$\mathbf{M}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{M}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

we have $\mathcal{W}_2^{2 \times 5} = \{\boldsymbol{\Pi}\boldsymbol{\Lambda}\mathbf{M}_1 \cup \boldsymbol{\Pi}\boldsymbol{\Lambda}\mathbf{M}_2 : \boldsymbol{\Pi} \in \mathcal{P}(2), \boldsymbol{\Lambda} \in \mathcal{D}(2)\}$. As another example, if $K = 5$, $P = 3$ and $Q(\mathbf{S}_1) < Q(\mathbf{S}_2) = Q(\mathbf{S}_3) = Q(\mathbf{S}_4) < Q(\mathbf{S}_5)$, then $P_m = 1$, $P_M = 4$, and with

$$\mathbf{M}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{M}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

$$\mathbf{M}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

we have $\mathcal{W}_3^{3 \times 5} = \{\boldsymbol{\Pi}\boldsymbol{\Lambda}\mathbf{M}_1 \cup \boldsymbol{\Pi}\boldsymbol{\Lambda}\mathbf{M}_2 \cup \boldsymbol{\Pi}\boldsymbol{\Lambda}\mathbf{M}_3 : \boldsymbol{\Pi} \in \mathcal{P}(3), \boldsymbol{\Lambda} \in \mathcal{D}(3)\}$.

As it will be shown further, Theorem 8 is very useful for showing the contrast property of a given criterion of the form given in the theorem; it suffices to prove Huber's superadditivity of the functional Q used in the criterion.

2.2 SHANNON'S ENTROPY CONTRAST

Shannon's entropy is a criterion from which it is known that contrast functions can be built. We briefly recall them here, and mention the theoretical arguments used to prove the contrast properties.

2.2.1 Simultaneous approach

The contrast property of

$$\mathcal{C}_h(\mathbf{B}) \doteq \log |\det \mathbf{B}| - \sum_{i=1}^K h(Y_i) , \quad (2.12)$$

$\mathbf{B} \in \mathbb{R}^{K \times K}$, can be proved by two simple ways. First, it is known that the opposite of $\text{KL}(\mathbf{Y}) = \sum_{i=1}^K h(Y_i) - h(\mathbf{Y})$ is a contrast function [Comon, 1994] (this is easily checked from Proposition 2 p. 16 and Corollary 3 p. 20). Hence, since $h(\mathbf{Y}) = h(\mathbf{X}) + \log |\det \mathbf{B}|$ from Eq. (1.74), $-\text{KL}(\mathbf{Y})$ equals $\log |\det \mathbf{B}| - \sum_{i=1}^K h(Y_i)$ up to a term not depending on \mathbf{B} , and $\mathcal{C}_h(\mathbf{B})$ benefits from the same contrast properties as $-\text{KL}(\mathbf{Y})$. A second simple way to prove the contrast properties of $\mathcal{C}_h(\mathbf{B})$ is to use Huber's superadditivity of $2^{h(\cdot)}$ resulting from the EPI theorem, Theorem 6 p. 38 (if Shannon's entropy is expressed in bits, or $e^{h(\cdot)}$ if the nat unit is chosen instead) ; indeed,

$$\mathcal{C}_h(\mathbf{B}) = \log |\det \mathbf{B}| - \sum_{i=1}^K \log(2^{h(Y_i)}) . \quad (2.13)$$

Obviously, all the non-mixing maximizers of the simultaneous contrast $\mathcal{C}_h(\mathbf{B})$ are global ones because the $K \times K$ non-mixing matrices are all PD-equivalent and because a S-BSS contrast is invariant under PD-equivalence preserving transforms. In the specific case where \mathbf{B} is constrained to be in $\mathcal{O}(K)$, the orthogonal version $\mathcal{C}_h^\perp(\mathbf{B})$ of $\mathcal{C}_h(\mathbf{B})$ simply reduces to $-\sum_{i=1}^K h(Y_i)$.

Remark 10 (From information theory to estimation theory) *Maximum likelihood is a technique in estimation theory for finding the optimal value of a model parameter Θ , in the sense that with this value of Θ , the obtained model makes the observations the most likely; with this value of the parameter, the probability that the outcome of the model yields the observed output is maximized. If $\mathbf{B} = \mathbf{A}^{-1}$, the probability of \mathbf{X} can be rewritten as (see Eq. (1.76)):*

$$p_{\mathbf{X}}(\mathbf{X}(t)) = |\det \mathbf{B}| \prod_{i=1}^K p_{S_i}(\mathbf{b}_i \mathbf{X}(t)) , \quad (2.14)$$

where $S_i(t) = \mathbf{b}_i \mathbf{X}(t)$: \mathbf{b}_i is the i -th row of \mathbf{B} . The likelihood of \mathbf{B} is given by

$$L(\mathbf{B}) = \prod_{t=1}^N \left(|\det \mathbf{B}| \prod_{i=1}^K p_{S_i}(\mathbf{b}_i \mathbf{X}(t)) \right) . \quad (2.15)$$

Since $\text{argmax}_{\mathbf{B}} L(\mathbf{B})$ is equal to $\text{argmax}_{\mathbf{B}} 1/N \log L(\mathbf{B})$, maximizing the likelihood is equivalent to maximizing

$$\log |\det \mathbf{B}| + 1/N \sum_{t=1}^N \sum_{i=1}^K \log p_{S_i}(Y_i(t)) . \quad (2.16)$$

This result is very close to the empirical counterpart $\hat{\mathcal{C}}_h(\mathbf{B})$ of $\mathcal{C}_h(\mathbf{B})$ as defined in Eq. (2.12) but where the theoretical expectation is replaced by the sample mean

$$\hat{\mathcal{C}}_h(\mathbf{B}) \doteq \log |\det \mathbf{B}| + 1/N \sum_{t=1}^N \sum_{i=1}^K \log p_{Y_i}(Y_i(t)) , \quad (2.17)$$

except that the argument of the logarithm is the marginal density of S_i evaluated at the given outcomes $\mathbf{b}_i \mathbf{X}$. Because the source densities are unknown, they have to be, in practice, either *a priori* assumed or guessed from the samples. If the criterion (2.16) is maximized by guessing at each step $p_{S_i} \leftarrow p_{Y_i}$, maximizing the likelihood is equivalent to minimizing the mutual information between the outputs. Hence, if the log-likelihood reaches its maximum value when $\mathbf{B} = \mathbf{A}^{-1}$ as it should, then

$$\frac{1}{N} \log L(\mathbf{B}) \simeq -\log |\det \mathbf{A}| - \sum_{i=1}^K h(S_i) . \quad (2.18)$$

In practice, when using the maximum likelihood estimation principle, the source densities have to be guessed. Consequently, a natural question is the following : “how rough can be the model of the source densities ?”. Actually, there exists a theoretical result which says that the maximum likelihood method is very robust to a departure of the source densities model (called target densities) from the true source densities. Except in some rare cases, it suffices to guess the sign of a non-polynomial moment of the source (i.e. if its density is sub- or super-Gaussian). More precisely, one can restrict the set of the target densities to only two well-chosen densities for the estimator to be locally consistent when one of them is used as the assumed source density. This is because whatever is the true density, these functions yield opposite sign for the non-polynomial moment [Hyvärinen, Karhunen, and Oja, 2001]. The choice between these functions can be done on-line, during the likelihood maximization.

To be complete, we point out without giving details that the maximum-likelihood method is equivalent to the Infomax approach [Bell and Sejnowski, 1995]; this was shown in [Cardoso, 1997].

2.2.2 Deflation approach

2.2.2.1 The contrast property

The deflation contrast property of Shannon's entropy is often badly understood using the *Central Limit Theorem* (CLT).

Basically, the CLT states that the distribution (i.e. the cumulative distribution function, cdf) of the sum of a collection of random variables converges to that of a Gaussian variable. Its simplest form deals with a sequence of i.i.d. random variables (see e.g. [Gray and Davisson, 2004]).

Theorem 9 (Central Limit Theorem (CLT)) Let X_1, \dots, X_K be a sequence of i.i.d. random variables with finite mean μ and variance σ^2 and common distribution $P_X(x)$. Then,

$$S_K = \frac{1}{\sqrt{K}} \sum_{i=1}^K (X_i - \mu)$$

converges in distribution¹ to a zero-mean Gaussian random variables with variance σ^2 .

In the above theorem, the convergence in distribution has to be well understood.

Definition 21 (convergence in distribution of S_K) Let P_S (resp. P_{S_K}) denote the cdf of S (resp. S_K). Then the random variable S_K “converges in distribution” to S if $\lim_{K \rightarrow \infty} P_{S_K}(x) = P_S(x)$ for all x where $P_S(x)$ is continuous.

Clearly, S_K in the above theorem resembles an arbitrary output Y_i as $Y_i = \sum_{j=1}^K W_{ij} S_j$, where the S_j are independent rv. However, we are facing a *weighted* sum, i.e. a sum of rv $W_{ij} S_j$ having possibly different variances. Moreover, a major limitation of the theorem is that the summed variables may not have different densities, but must necessarily share the same cdf. This requirement is not really necessary as an extended form of the CLT (due to Lindeberg in 1922, see [Feller, 1966, Rényi, 1966, Cramér, 1946]) allows one to deal with summed variables having arbitrary densities (as well as arbitrary variances, i.e. arbitrary mixture weights). Hence, the CLT matches our mixture model.

In our BSS context, it is thought that the CLT tells us that minimizing Shannon’s entropy (i.e. make the output “different” from a Gaussian, where entropy is used as a ‘the non-Gaussianity index) gives the original, unmixed source signals (still under a fixed-variance constraint). However, this intuitive reasoning does not constitute an absolute proof because the CLT is a limit theorem. To our knowledge, there is no formal proof that, whatever the non-Gaussianity index (which is a concept that is not clearly defined), the finite number of samples and the finite number K of sources (most often relatively few), the index will decrease until a satisfactory solution is found.

Nevertheless, the suitable use of Shannon’s entropy for BSS can be proved using the EPI which is definitely and fortunately not a limit theorem, under fixed variance that is, under fixed norm for $\|\mathbf{w}\|$ because

$$\text{Var}[\mathbf{w}S] = \sum_{i=1}^K w_i^2 \text{Var}[S_i] \stackrel{\mathcal{A}_7}{=} \|\mathbf{w}\|^2 . \quad (2.19)$$

Indeed, remind that the following result can be proved in Lemma 5 (p. 38): if $h(\mathbf{w}S)$ reaches its minimum value, then $\mathbf{w}X \propto S_j$ for $j \in \{\underset{k}{\operatorname{argmin}} h(S_k)\}$.

Let us define the following criterion

$$\begin{aligned} \mathcal{C}_h(\mathbf{b}_i) &\doteq \frac{1}{2} \log \text{Var}[\mathbf{b}_i X] - h(\mathbf{b}_i X) \\ &= -h\left(\frac{Y_i}{\sqrt{\text{Var}[Y_i]}}\right) , \end{aligned} \quad (2.20)$$

¹Observe that we can say that the distribution (resp. the characteristic function) of the normalized sum converges to the distribution (resp. characteristic function) of a Gaussian rv, but this may not be true for the pdf, as the summed rv might be discrete.

whose maximization, with $\mathbf{w}_i = \mathbf{b}_i \mathbf{A}$ and thus $\mathbf{Y}_i = \mathbf{w}_i \mathbf{S}$ is equivalent to that of

$$\begin{aligned}\tilde{\mathcal{C}}_h(\mathbf{w}_i) &\doteq \mathcal{C}_h(\mathbf{b}_i) \\ &= -h\left(\frac{\mathbf{w}_i \mathbf{S}}{\sqrt{\text{Var}[\mathbf{w}_i \mathbf{S}]}}\right).\end{aligned}\quad (2.21)$$

Observe in passing that, from Eq. (2.19), $\tilde{\mathcal{C}}_h(\mathbf{w}_i) = -h\left(\frac{\mathbf{w}_i}{\|\mathbf{w}_i\|} \mathbf{S}\right)$ and does not depend on the magnitude of the transfer vector \mathbf{w}_i .

Then, in order to prove that $\mathcal{C}_h(\mathbf{b}_i)$ is a D-BSS contrast, it is sufficient to prove the associated contrast property of $\tilde{\mathcal{C}}_h(\mathbf{w}_i)$. This is clearly the case as stated by Lemma 5 and the following corollary, which results from Eq. (2.19).

Corollary 7 *Minimizing $h(\mathbf{w}\mathbf{S})$ subject to $\mathbf{w} \in \mathcal{S}(K)$ is equivalent to minimizing the unconstrained entropy $h\left(\mathbf{w}\mathbf{S}/\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]}\right)$ since both approaches consist in minimizing the entropy of a unit-variance output with respect to the mixture weights.*

More concretely, we have proven the following theorem (see the related paper of [Vrins et al., 2007b]).

Theorem 10 (Global maximum of \mathcal{C}_h , deflation approach) *Suppose that*

$$h(\mathbf{S}_1) = \dots = h(\mathbf{S}_k) < h(\mathbf{S}_{k+1}) \leq \dots \leq h(\mathbf{S}_K).$$

Then,

$$\underset{\mathbf{w} \text{ s.t. } \|\mathbf{w}\| = \lambda}{\operatorname{argmax}} \tilde{\mathcal{C}}_h(\mathbf{w}) = \{\pm \lambda \cdot \mathbf{e}_1, \dots, \pm \lambda \cdot \mathbf{e}_k\}. \quad (2.22)$$

Note that the minimization of $h\left(\mathbf{Y}_i/\sqrt{\text{Var}[\mathbf{Y}_i]}\right)$ is equivalent to the maximization of the negentropy index $h[\phi_{\mathbf{Y}_i}] - h(\mathbf{Y}_i)$; denoting by ϕ_X the density of a centered rv with Gaussian density with same variance σ^2 of X , i.e.

$$\phi_X(x) \doteq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}, \quad (2.23)$$

then

$$\begin{aligned}h[\phi_{\mathbf{Y}_i}] - h(\mathbf{Y}_i) &= \frac{1}{2} \log(2\pi e) + \frac{1}{2} \log \text{Var}[\mathbf{Y}_i] \\ &\quad - \left(h\left(\frac{\mathbf{Y}_i}{\sqrt{\text{Var}[\mathbf{Y}_i]}}\right) + \frac{1}{2} \log \text{Var}[\mathbf{Y}_i] \right) \\ &= -h\left(\frac{\mathbf{Y}_i}{\sqrt{\text{Var}[\mathbf{Y}_i]}}\right) + \text{cst}.\end{aligned}\quad (2.24)$$

Let us now turn to the non-mixing local maxima of $\mathcal{C}_h(\mathbf{w})$, under the $\mathbf{w} \in \mathcal{S}(K)$ constraint.

Remark 11 A natural question is the following: “what are the connections between Shannon’s entropy (and more specifically, the EPI Theorem) and the CLT?”. Answering this question is not an easy task, actually. Shannon’s entropy is known to be a non-Gaussianity index in the sense that the entropy of a random variable is upper bounded by the entropy of a Gaussian variable with the same variance, see Theorem 5 p. 34 (just like the absolute or square kurtosis reach their minimum value for Gaussian functions under a fixed-variance constraint). Sometimes, it is said that the “essence of the EPI is to express that the sum of independent variables tends to be more Gaussian than each of the individual components” (see e.g. [Verdu and Guo, 2006]); if this reveals to be true, here is the sought connection: the EPI would be a non-asymptotic form of the CLT, in the sense that it holds for a finite number of variables. Unfortunately, we believe that this viewpoint should be explained in more detail: the entropy is sensitive to the variance and for independent random variables, the variance of a sum is the sum of the variances of the random variables involved in the sum; the “increase of entropy” corresponding to the difference $2^{2h(\sum_{i=1}^K S_i)} - \sum_{i=1}^K 2^{2h(S_i)}$ could thus either result from a “Gaussianization” of the density shape of $\sum_{i=1}^K S_i$ compared to those of the individual sources (which correspond to the essence of the CLT), but also from the variance increase, or both. Letting $S = \sum_{i=1}^K S_i$, is the entropy of S larger than each of the individual entropies because the entropy of the unit-variance random variable $S/\sqrt{\text{Var}[S]}$ increases, because for all i $\text{Var}[S] > \text{Var}[S_i]$ or because of a joint effect? Only the occurrence of the first situation would express a connection between EPI and CLT: entropy is a non-Gaussianity index under a same-variance constraint. For instance comparing the entropy of two variables with different variances does not tell anything about how close their distribution functions are from the Gaussian cdf, it is simply a non-sense!

We know from the CLT that (with a slight abuse of notation that has the advantage to be illustrative)

$$\lim_{\sharp[I(\mathbf{w})] \rightarrow \infty} h(wS) = \frac{1}{2} \log 2\pi e . \quad (2.25)$$

This result is not so strong as it is again a limit result in terms of (here) the number of sources. The hot question is the following: is this convergence monotonic in $\sharp[I(\mathbf{w})]$? This would solve a long-standing conjecture! In the simplest case, the answer has recently been proved to be positive: the entropy of $1/K \sum_{i=1}^K S_i$ (where the S_i are i.i.d.) is a non-decreasing sequence for every K [Madiman and Barron, 2006, Artstein et al., 2004]. More specifically:

$$h\left(\frac{S_1 + \dots + S_K}{\sqrt{K}}\right) \geq h\left(\frac{S_1 + \dots + S_{K-1}}{\sqrt{K-1}}\right) . \quad (2.26)$$

This key result would clearly make sense to the naïve CLT-based justification to Shannon’s entropy contrast for deflation provided that it extends to non i.i.d. variables (there exists a non-i.i.d. version of the above result, but it is more complicated to interpret). Note that a less general result (but more

general than the CLT except about the i.i.d. assumption) showed that if the S_i are i.i.d. with pdf p_S (and assume unit-variance to simplify the notations) $\text{KL}(\mathbf{w}\mathbf{S}, \mathbf{w}\tilde{\mathbf{S}}) \leq \text{KL}[p_S\|\phi]$ where $\tilde{\mathbf{S}}$ is a vector with independent Gaussian standardized components. The inequality is strict unless $p_S \equiv \phi$ almost everywhere or $\sharp[I(\mathbf{w})] = 1$. This means indeed that mixing Gaussianizes: if all the sources share the same pdf, a given output is more Gaussian than a source in the KL-sense, and even for finite K (this is the improvement compared to the CLT) [Zamir and Feder, 1993].

For more details about the connections between the EPI and the CLT, we refer the reader to the papers [Zamir and Feder, 1993, Madiman and Barron, 2006, Barron, 1984, 1986, Artstein et al., 2004]

2.2.2.2 Non-mixing local maxima

In subsection 2.2.2.1, it was shown that the Shannon entropy-based criteria for BSS is i) not sensitive to scaling and ii) reaches its global maximum point if and only if the lowest entropic source has been recovered. But at this step, nothing is known about the possible existence of non-mixing maximum points.

In order to check if a unit-norm vector $\mathbf{w} = \mathbf{e}_i$ is a local maximum point of the entropy, we shall analyze a second order development of Shannon's entropy. We set $\|\mathbf{w}\| = 1$ for convenience. The starting point is an expansion up to second order of the entropy of a random variable Y slightly contaminated with another variable δY , possibly dependent from Y , which has been established in [Pham, 2005]:

$$h(Y + \delta Y) \approx h(Y) + E[\psi_Y(Y)\delta Y] + \frac{E[\text{Var}[\delta Y|Y]\psi'_Y(Y) - (E[\delta Y|Y])'^2]}{2}. \quad (2.27)$$

In this equation, ψ_Y is the *score function* of Y , defined as²

$$\psi_Y(Y) \doteq -(\log p_Y(Y))' = -\frac{p'_Y(Y)}{p_Y(Y)}, \quad (2.28)$$

and p_Y is the pdf of Y , $'$ denotes the derivative (here, with respect to Y), and $E[\cdot|Y]$ and $\text{Var}[\cdot|Y] = E[\cdot^2|Y] - E^2[\cdot|Y]$ denote the conditional expectation and conditional variance given Y , respectively. The score function satisfies the following.

Property 4 For well behaved densities, the score is a zero-mean function, satisfying $E[X\psi_X] = 1$ and $E[\psi'_X] = E[\psi_X^2]$. Furthermore, the inequality $E[\psi_X^2]\text{Var}[X] \geq 1$ holds with equality if and only if X is Gaussian.

This proposition is proved in Section 2.6.2, p. 84.

²In this work, we use the score function definition presented in [Pham, 2002]. However, several authors define this function with the opposite sign. The reader should keep this difference in mind.

Theorem 10 (p. 55) tells us that $\mathcal{C}_h(\mathbf{b})$ is a D-BSS contrast. Based on the next theorem, Corollary 9 states the complete extraction property of the contrast.

Theorem 11 (Subset of local maximum point of $\tilde{\mathcal{C}}_h(\mathbf{w})$) *The constrained entropy $h(\mathbf{w}\mathbf{S})$ s.t. $\mathbf{w} \in \mathcal{S}(K)$ reaches a local minimum at $\mathbf{w} = \pm\mathbf{e}_j$, $j \in \{1, \dots, K\}$, the j -th row of the $K \times K$ identity matrix, if \mathbf{S}_j is non-Gaussian, or a global maximum otherwise. In other words, the criterion $\tilde{\mathcal{C}}_h(\mathbf{w})$ subject to the $\|\mathbf{w}\| = \lambda$ constraint is stationary when $\mathbf{w} \in \{\pm\lambda\mathbf{e}_1, \dots, \pm\lambda\mathbf{e}_K\}$. Further, if $\widetilde{\text{argmax}}(\cdot)$ (resp. $\widetilde{\text{argmin}}(\cdot)$) denotes the set of local maximum (resp. minimum) points of \cdot and \mathcal{I}_G denotes the set of indexes of the Gaussian sources,*

$$\widetilde{\text{argmax}}_{\mathbf{w} \text{ s.t. } \|\mathbf{w}\| = \lambda} \tilde{\mathcal{C}}_h(\mathbf{w}) \supseteq \{\pm\lambda\mathbf{e}_i : i \in \{1, \dots, K\} \setminus \mathcal{I}_G\}, \quad (2.29)$$

and

$$\widetilde{\text{argmin}}_{\mathbf{w} \text{ s.t. } \|\mathbf{w}\| = \lambda} \tilde{\mathcal{C}}_h(\mathbf{w}) = \{\pm\lambda\mathbf{e}_i : i \in \mathcal{I}_G\}. \quad (2.30)$$

The proof of this Theorem is given in the Appendix of the Chapter, in Section 2.6.3 (p. 86). Note that, by Theorem 10, the global maximum of $\mathcal{C}_h(\mathbf{b})$ s.t. $\|\mathbf{w}\| = \lambda$ is reached at $\pm\lambda\mathbf{e}_k$ where $k \in \text{argmin}_i h(\mathbf{S}_i)$.

2.2.3 Partial approach

Proving the partial contrast property of

$$\mathcal{C}_h(\mathbf{B}) \doteq \log |\det(\mathbf{B}\Sigma_X\mathbf{B}^T)| - \sum_{i=1}^P h(Y_i), \quad \mathbf{B} \in \mathbb{R}^{P \times K} \quad (2.31)$$

is immediate. This results from the superadditivity of $Q = 2^{h(\cdot)}$, which is a consequence of the EPI given in Section 2.2.1. The notation of the entropic partial contrast is the same as the simultaneous one because they are identical when $P = K$; in this case, maximizing $\log |\det \mathbf{B}| - \sum_{i=1}^K h(Y_i)$ is equivalent to maximizing $\log \det(\mathbf{B}\Sigma_X\mathbf{B}^T) - \sum_{i=1}^K h(Y_i)$ with respect to \mathbf{B} since $\frac{1}{2} \log |\det(\mathbf{B}\Sigma_X\mathbf{B}^T)| = \log |\det \mathbf{B}| + \text{cst}$. A recent result [Pham, 2006b], recalled in the next theorem, states the stationarity of

$$\tilde{\mathcal{C}}_h(\mathbf{W}) \doteq \log \det(\mathbf{W}\mathbf{W}^T) - \sum_{i=1}^P h(Y_i), \quad \mathbf{W} = \mathbf{B}\mathbf{A} \in \mathbb{R}^{P \times K} \quad (2.32)$$

at non-mixing points.

Theorem 12 (Subset of stationary points of $\tilde{\mathcal{C}}_h(\mathbf{W})$) *The non-mixing matrices $\mathbf{W} \in \mathcal{W}^{P \times K}$ are stationary points of $\tilde{\mathcal{C}}_h(\mathbf{W})$. More precisely, these matrices \mathbf{W} are local maximum points of the criterion if none of the $\mathbf{w}_i\mathbf{S}$ (that are proportional to distinct sources) is Gaussian.*

To see how this result extends to the non-mixing points of $\mathcal{C}_h(\mathbf{B})$, observe that both criteria are equal up to a possible constant term:

$$\begin{aligned}\log |\det(\mathbf{B}\Sigma_X\mathbf{B}^T)| &= \log |\det(\mathbf{B}\mathbf{A}\Sigma_S(\mathbf{B}\mathbf{A})^T)| \\ &= \log(\det(\mathbf{W}\mathbf{W}^T)\det\Sigma_S) \\ &= \log\det(\mathbf{W}\mathbf{W}^T) + \sum_{i=1}^K \log\text{Var}[S_i] .\end{aligned}\quad (2.33)$$

The last equality results from the diagonal form of matrix Σ_S (equal to the identity matrix if the independent sources have unit variances). Note that the absolute values vanish as both $\mathbf{W}\mathbf{W}^T \succeq 0$ and $\Sigma_S \succeq 0$. In particular, we have

$$\widetilde{\arg\max}_{\mathbf{B} \in \mathcal{M}^{P \times K}} \mathcal{C}_h(\mathbf{B}) = \widetilde{\arg\max}_{\mathbf{B}: \mathbf{B}\mathbf{A} \in \mathcal{M}^{P \times K}} \tilde{\mathcal{C}}_h(\mathbf{B}\mathbf{A}) \quad (2.34)$$

Hence, we have the following corollary.

Corollary 8 (Subset of stationary points of $\mathcal{C}_h(\mathbf{B})$) *The demixing matrices $\mathbf{B} \sim_u \mathbf{A}^{-1}$ are stationary points of $\mathcal{C}_h(\mathbf{B})$. More precisely, these matrices \mathbf{B} are local maximum points of the criterion if none of the $\mathbf{b}_i\mathbf{X}$ (that are proportional to distinct sources since $\mathbf{B} \sim_u \mathbf{A}^{-1}$) is Gaussian.*

Remark 12 *The first result suggesting the use of the opposite of the sum of the marginal entropies of $P \leq K$ outputs for the extraction of $P \leq K$ signals can be found in [Cruces et al., 2001]. However, in this specific case, the contrast is orthogonal as the mixing matrix is supposed to be orthogonal, and hence the demixing matrix \mathbf{B} is forced to be semi-orthogonal, that is $\mathbf{B}\mathbf{B}^P = \mathbf{I}_P$ implying that so is \mathbf{W} (see Remark 4 p. 48). This is also proved by the same authors in [Cruces et al., 2004] by using the negentropy instead of the entropy. Negentropy always satisfies the positivity requirement with equality if and only if the output is Gaussian and the weak form of convexity results from the EPI. The result obtained through Pham's approach is more general in the sense that the constraint is included in the criterion.*

2.3 MINIMUM RANGE CONTRAST

Shannon's entropy was seen to be the extended Rényi's entropy (ERE) with $r = 1$. Another remarkable case of the ERE is $h_{0,\Omega}$ (see Eq. (1.109)). This section aims at analyzing the contrast properties of this criterion.

2.3.1 Support and Brunn-Minkowski Inequality

Clearly, if $Q(X) = \mu[\Omega(X)]$, then

$$Q(\alpha X) = |\alpha| \mu[\Omega(X)] , \quad (2.35)$$

which shows that the first requirement of Huber's superadditivity given in Theorem 8 p. 50 is fulfilled. To prove the second requirement, we consider the so-called *Brunn-Minkowski Inequality* (BMI) [Gardner, 2002].

Theorem 13 (Brunn-Minkowski Inequality (BMI)) *If \mathcal{X} and \mathcal{Y} are two compact convex sets with nonempty interiors (i.e. measurable) in \mathbb{R}^K , then for any $\alpha, \beta > 0$:*

$$\text{Vol}^{1/K}[\alpha\mathcal{X} + \beta\mathcal{Y}] \geq \alpha\text{Vol}^{1/K}[\mathcal{X}] + \beta\text{Vol}^{1/K}[\mathcal{Y}] . \quad (2.36)$$

The operator $\text{Vol}[\cdot]$ stands for volume. The operator “+” is defined on sets as $\mathcal{X} + \mathcal{Y} = \{x + y : x \in \mathcal{X}, y \in \mathcal{Y}\}$. The equality holds when \mathcal{X} and \mathcal{Y} are equal up to translation and dilatation (i.e. when they are homothetic).

In 1990, Dembo gave a simultaneous proof of the EPI and BMI theorems [Dembo, 1990].

We use here one-dimensional sets (the support of one-dimensional signals) and the Lebesgue measure $\mu[\cdot]$ as the volume $\text{Vol}[\cdot]$ operator.

Inequality (2.36) has been extended in [Costa and Cover, 1984, Cover and Thomas, 1991] to non-convex bodies; in this case however, to the author's knowledge, the *strict equality* and *strict inequality* cases were not discussed in the literature. Therefore, the following lemma, which is an extension of the BMI theorem in the specific $K = 1$ case, states the conditions for the strict equality to hold. A restricted form of this lemma appeared in [Vrins et al., 2006].

Lemma 6 (Extended BMI) *Consider two independent bounded random variables X and Y . Suppose that $\mu[\Omega(X)] > 0$, $\mu[\Omega(Y)] > 0$, with $\Omega(X) \subset \mathbb{R}$, $\Omega(Y) \subset \mathbb{R}$. Then:*

$$\mu[\Omega(X + Y)] \geq \mu[\Omega(X)] + \mu[\Omega(Y)] ,$$

with equality if and only if $\mu[\overline{\Omega}(X) \setminus \Omega(X)] = \mu[\overline{\Omega}(Y) \setminus \Omega(Y)] = 0$, where $\overline{\Omega}(\cdot)$ denotes the convex hull of $\Omega(\cdot)$, that is, the smallest interval including the one-dimensional support $\Omega(\cdot)$.

The proof is given at the end of the Chapter, in Section 2.6.4 (p. 87). Clearly, since the support measure is a positive quantity and since the second power is a monotonously increasing mapping, the above lemma states the second requirement of Huber's superadditivity given in Theorem 8 (p. 50).

2.3.2 Properties of the range

The range is a specific case of the support measure. The range $R(X)$ of a random variable X is the measure of the convex hull of $\Omega(X)$:

$$R(X) \doteq \mu[\overline{\Omega}(X)] , \quad (2.37)$$

implying $R(\alpha X) = |\alpha|R(X)$.

Then, considering the range criterion instead of the support is exactly the same as working with the support if the source supports are convex sets. Actually, the range possesses a stronger property. Assume that X and Y are two independent random variables, then, from the Extended BMI lemma:

$$R(X + Y) = R(X) + R(Y) . \quad (2.38)$$

The last result can also be seen as a consequence of the fact that p_{X+Y} is the convolution of p_X and p_Y [Hirschman and Widder, 1955, Feller, 1966].

□

Because both $R(X) > 0$ and $R(Y) > 0$, Eq. (2.38) implies the strict superadditivity

$$R^2(X + Y) > R^2(X) + R^2(Y) \quad (2.39)$$

for any pair of independent bounded random variables X and Y . This results from the strict equality in the Extended BMI lemma (Lemma 6).

Hence, one has:

$$R(\mathbf{b}_i \mathbf{X}) = R(\mathbf{b}_i \mathbf{A} \mathbf{S}) = \sum_{j=1}^K |W_{ij}| R(S_j) . \quad (2.40)$$

The above properties will be useful for proving the contrast properties of range-based criteria in the three extraction schemes (simultaneous, deflation and partial separation).

2.3.3 Simultaneous approach

The minimum range approach for the simultaneous extraction of bounded sources has been first introduced in [Pham, 2000]. The following criterion $\mathcal{C}_R(\mathbf{B})$ was proposed:

$$\mathcal{C}_R(\mathbf{B}) \doteq \log |\det \mathbf{B}| - \sum_{i=1}^K \log R(\mathbf{b}_i \mathbf{X}), \quad (2.41)$$

which has the same form as the one given in Theorem 8 (p. 50) with $Q(\cdot) = R(\cdot)$.

As usual, this criterion has to be maximized with respect to the demixing matrix \mathbf{B} .

It has been shown in [Pham, 2000] that $\mathcal{C}_R(\mathbf{B})$ is a S-BSS contrast. Note that the proof is trivial using the superadditivity of the range combined with Theorem 8.

Similarly to the Shannon entropy-based criterion with $\mathbf{W} = \mathbf{B}\mathbf{A}$, maximizing $\mathcal{C}_R(\mathbf{B})$ over the set $\mathcal{M}(K)$ of $K \times K$ non-singular matrices is equivalent to maximizing

$$\tilde{\mathcal{C}}_R(\mathbf{W}) \doteq \log |\det \mathbf{W}| - \sum_{i=1}^K \log \left[\sum_{j=1}^K |W_{ij}| R(S_j) \right] , \quad (2.42)$$

also over $\mathcal{M}(K)$ because of Eq. (2.40) and $\log |\det \mathbf{B}| = \log |\det \mathbf{W}| - \log |\det \mathbf{A}|$. In particular,

$$\widetilde{\operatorname{argmax}}_{\mathbf{B} \in \mathcal{M}(K)} \mathcal{C}_R(\mathbf{B}) = \widetilde{\operatorname{argmax}}_{\mathbf{B}: \mathbf{BA} \in \mathcal{M}(K)} \tilde{\mathcal{C}}_R(\mathbf{BA}) . \quad (2.43)$$

A point \mathbf{B} maximizing \mathcal{C}_R is related to a given point \mathbf{W} maximizing $\tilde{\mathcal{C}}_R$ by the relation $\mathbf{W} = \mathbf{BA}$.

2.3.4 Deflation approach

We present here our results showing that $-R(Y_i)$ can be used as a deflation contrast if $\operatorname{Var}[Y_i]$ is kept constant (these results can be found in [Vrins et al., 2007a]). However, we would like to point out that another proof, based on information theory, has been provided simultaneously and independently in [Cruces and Duran, 2004] (see the related comment in Section 2.3.6).

Let us first observe that $R(\mathbf{b}_i \mathbf{X}) = R(\mathbf{w}_i \mathbf{S})$. Then, the range of the fixed variance i -th output equals $R\left(\mathbf{w}_i \mathbf{S} / \sqrt{\operatorname{Var}[Y_i]}\right) = R(\mathbf{w}_i \mathbf{S}) / \sqrt{\operatorname{Var}[Y_i]}$, which does not depend on the magnitude of Y_i . Based on the above contrast forms, we define the following criterion:

$$\mathcal{C}_R(\mathbf{b}_i) \doteq -R\left(\mathbf{b}_i \mathbf{X} / \sqrt{\operatorname{Var}[\mathbf{b}_i \mathbf{X}]}\right) . \quad (2.44)$$

Clearly, maximizing $\mathcal{C}_R(\mathbf{b}_i)$ with respect to \mathbf{b}_i is equivalent to maximizing

$$\tilde{\mathcal{C}}_R(\mathbf{w}_i) \doteq -R\left(\mathbf{w}_i \mathbf{S} / \sqrt{\operatorname{Var}[\mathbf{w}_i \mathbf{S}]}\right) , \quad (2.45)$$

where $\mathbf{w}_i = \mathbf{b}_i \mathbf{A}$.

Note that the above denominators, that are equal to $\sqrt{\operatorname{Var}[Y_i]}$, can be omitted if \mathbf{w}_i is constrained to have a fixed norm because of Eq. (2.19). In the following, we omit the index of \mathbf{w} as it does not matter if we focus on an arbitrary output $\mathbf{Y} = \mathbf{b} \mathbf{X} = \mathbf{w} \mathbf{S}$. Further, since $R(\mathbf{Y})$ is not sensitive to the sign of the elements of \mathbf{b} , we can freely assume $\mathbf{w} \in \mathcal{V}_K^\lambda$ where

$$\mathcal{V}_K^\lambda \doteq \{\mathbf{w} \in \mathbb{R}^K \text{ s.t. } \|\mathbf{w}\| = \lambda, \mathbf{w}(j) > 0 \forall 1 \leq j \leq K\} \quad (2.46)$$

is nothing but the intersection of \mathbb{R}_+^K with the centered K -dimensional sphere of radius λ . Consider now the following theorem.

Theorem 14 (Global maximum of \mathcal{C}_R , deflation approach) *Suppose that*

$$R(\mathbf{S}_1) = \dots = R(\mathbf{S}_k) < R(\mathbf{S}_{k+1}) \leq \dots \leq R(\mathbf{S}_K) . \quad (2.47)$$

Then, for any vector \mathbf{w} in \mathcal{V}_K^λ , one gets

$$\operatorname{argmax}_{\mathbf{w} \in \mathcal{V}_K^\lambda} \tilde{\mathcal{C}}_R(\mathbf{w}) = \{\lambda \cdot \mathbf{e}_1, \dots, \lambda \cdot \mathbf{e}_k\} .$$

Proof: The proof is very similar to that of the Shannon entropy case by setting $2^h \rightarrow R$ and using the BMI instead of the EPI (see pp. 38 and 39). Recall that $I(\mathbf{w})$ is the vector containing the position indexes of the non-zero entries of \mathbf{w} (Eq. (1.96)); assume that $\mathbf{w}\mathbf{S}$ is not proportional to a source, i.e. $\#[I(\mathbf{w})] \geq 2$. By the extended BMI, we have

$$\begin{aligned} R^2(\mathbf{w}\mathbf{S}) &\geq \sum_{i \in I(\mathbf{w})} w_i^2 R^2(\mathbf{S}_i) \\ &\stackrel{(a)}{=} R^2(\mathbf{S}_{k'}) + \underbrace{\sum_{i \in I(\mathbf{w}) \setminus \{k'\}} w_i^2 (R^2(\mathbf{S}_i) - R^2(\mathbf{S}_{k'}))}_{\geq 0} \\ &\geq R^2(\mathbf{S}_p) \end{aligned} \quad (2.48)$$

where $k' \in \operatorname{argmin}_{i \in I(\mathbf{w})} R(\mathbf{S}_i)$ and $p \in \{1, \dots, k\}$.

In the above chain of inequalities, (a) results from the $\|\mathbf{w}\| = 1$ constraint (which is equivalent to assume $\operatorname{Var}[\mathbf{S}_i] = \operatorname{Var}[\mathbf{w}\mathbf{S}] = 1$). On the other hand, it is obvious that in the $\#[I(\mathbf{w})] = 1$ case, the strict inequality holds except if $I(\mathbf{w}) \in \{1, \dots, k\}$. Hence, the strict equality case occurs if and only if $I(\mathbf{w}) \in \{1, \dots, k\}$ (implying $\#[I(\mathbf{w})] = 1$), i.e. when $\mathbf{w}\mathbf{S} \propto \mathbf{S}_i$, $i \in \{1, \dots, k\}$.

□

Another proof is given in the Appendix at the end of the Chapter, in Section 2.6.5 (p. 89).

The above theorem guarantees that $\tilde{\mathcal{C}}_R(\mathbf{w})$ and $\mathcal{C}_R(\mathbf{b})$ reach their global maximum point when and only when one of the sources with the lowest range has been extracted. Because of the scale invariance, one can set $\lambda = 1$ in the analysis of $\tilde{\mathcal{C}}_R(\mathbf{w}\mathbf{S})$ even though the mathematical developments can easily be extended to other values of λ .

Theorem 15 (Subset of local maxima of \mathcal{C}_R , deflation approach) *The function $\tilde{\mathcal{C}}_R(\mathbf{w})$, subject to $\mathbf{w} \in \mathcal{V}_K^1$, admits a local maximum for $\mathbf{w} = \mathbf{e}_i$, $1 \leq i \leq K$. In other words,*

$$\widetilde{\operatorname{argmax}}_{\mathbf{w} \in \mathcal{V}_K^1} \tilde{\mathcal{C}}_R(\mathbf{w}) \supseteq \{\mathbf{e}_i : i \in \{1, \dots, K\}\} . \quad (2.49)$$

Sketch of proof: Consider two vectors $\mathbf{p} \in \mathcal{V}_K^1$, $\mathbf{q} \in \mathcal{V}_K^1$, and let us introduce the associate contrast difference $\Delta\tilde{\mathcal{C}}_R(\mathbf{p}, \mathbf{q})$ defined as:

$$\Delta\tilde{\mathcal{C}}_R(\mathbf{p}, \mathbf{q}) \doteq \tilde{\mathcal{C}}_R(\mathbf{p}) - \tilde{\mathcal{C}}_R(\mathbf{q}) . \quad (2.50)$$

The proof shows that for any $\hat{\mathbf{e}}_i \in \mathcal{V}_K^1$ sufficiently close to (but different from) \mathbf{e}_i , we have $\Delta\tilde{\mathcal{C}}_R(\mathbf{e}_i, \hat{\mathbf{e}}_i) > 0$. The detailed proof is given in the Appendix at the end of the Chapter, in Section 2.6.6 (p. 90).

Corollary 9 (Complete extraction) *Assuming that the first $p - 1$ sources have already been extracted; then, the global maximum of $\tilde{\mathcal{C}}_R(\mathbf{w}_p)$ subject to $\mathbf{w}_p \in \mathcal{V}_K^1$ and $\mathbf{w}_p \mathbf{w}_r^T = 0$ for all $1 \leq r < p$ is obtained for $\mathbf{w}_p \in \{\mathbf{e}_i : \tilde{\mathcal{C}}_R(\mathbf{e}_i) = \tilde{\mathcal{C}}_R(\mathbf{e}_p)\}$.*

By Theorem 15, we know that $\tilde{\mathcal{C}}_R(\mathbf{w})$ s.t. $\mathbf{w} \in \mathcal{V}_K^1$ reaches a local maximum if $\mathbf{w} \in \{\mathbf{e}_1, \dots, \mathbf{e}_K\}$. Then, assuming that the first $p - 1$ sources have already been extracted, a p -th source can be found by updating \mathbf{w}_p where $\mathbf{w}_p(1) = \dots = \mathbf{w}_p(p-1) = 0$. Next, discarding the first $p - 1$ sources and setting $K \leftarrow K - p + 1$, Theorem 14 is used to prove that the global maximum of $\tilde{\mathcal{C}}_R(\mathbf{w})$, $\mathbf{w} \in \mathcal{V}_K^1$ equals now $\tilde{\mathcal{C}}_R(\mathbf{e}_p)$ and is reached for $\mathbf{w} \in \{\mathbf{e}_i : \tilde{\mathcal{C}}_R(\mathbf{e}_i) = \tilde{\mathcal{C}}_R(\mathbf{e}_p)\}$, $p \leq i \leq K$.

2.3.5 Partial approach

This case is exactly similar to that of the simultaneous contrast, and from the second claim of Theorem 8 (p. 50) with $Q(\cdot) = R(\cdot)$, we define the following contrast over $\mathcal{M}^{P \times K}$ (the set of $P \times K$ matrices with row-rank equal to $P \leq K$)

$$\mathcal{C}_R(\mathbf{B}) \doteq \log \det(\mathbf{B} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{B})^T - \sum_{i=1}^P \log R(\mathbf{b}_i \mathbf{X}) , \quad (2.51)$$

whose maximization is equivalent to the maximization of

$$\tilde{\mathcal{C}}_R(\mathbf{W}) \doteq \log \det(\mathbf{W} \mathbf{W}^T) - \sum_{i=1}^P \log \left[\sum_{j=1}^K |W_{ij}| R(\mathbf{s}_j) \right] \quad (2.52)$$

over the same subset of $\mathbb{R}^{P \times K}$ (see Section 2.2.3 and Eq. (2.40)). In particular,

$$\widetilde{\arg\max}_{\mathbf{B} \in \mathcal{M}^{P \times K}} \mathcal{C}_R(\mathbf{B}) = \widetilde{\arg\max}_{\mathbf{B}: \mathbf{B}\mathbf{A} \in \mathcal{M}^{P \times K}} \tilde{\mathcal{C}}_R(\mathbf{B}\mathbf{A}) . \quad (2.53)$$

Clearly, the contrast property combined to Corollary 6 (p. 50) with $Q(\cdot) = R(\cdot)$ tells us that the two above criteria reach their global maximum points if and only if $\mathbf{W} = \mathbf{B}\mathbf{A} \in \mathcal{W}_P^{P \times K}$, where the last set is defined as in Corollary 6. However, there is no information regarding the possible non-mixing *local* maxima of the criterion, i.e. the local maximum points corresponding to $\mathbf{W} = \mathbf{B}\mathbf{A} \in \mathcal{W}^{P \times K}$ (and not only in $\mathcal{W}_P^{P \times K} \subset \mathcal{W}^{P \times K}$).

Even though it was true for $Q(\cdot) = 2^{h(\cdot)}$, we do not claim that, generally speaking, $f(\mathbf{B})$, as given in Theorem 8 (p. 50), is locally maximized once $\mathbf{B}\mathbf{A} \in \mathcal{W}^{P \times K}$, even under the class II superadditivity assumption on functional Q . Nevertheless, this result holds true for the $Q(\cdot) = R(\cdot)$ case, just as for Shannon's entropy power. This is indicated by the following theorem [Vrins and Pham, 2007], proved in the Appendix of the Chapter, in Section 2.6.7 (p. 91).

Theorem 16 (Non-mixing matrices are local maximum points of \mathcal{C}_R) *The criterion $\mathcal{C}_R(\mathbf{B})$ admits a local maximum at any point \mathbf{B} for which $\mathbf{B}\mathbf{A} \in \mathcal{W}^{P \times K}$ (i.e. $\mathbf{B} \sim_u \mathbf{A}^{-1}$).*

Consequently, $\mathcal{C}_R(\mathbf{B})$ reaches a global maximum if and only if $\mathbf{BA} \in \mathcal{W}_P^{P \times K}$ (Theorem 8 p. 50 and Corollary 6 p. 50) and a local maximum point if $\mathbf{BA} \in \mathcal{W}^{P \times K}$ (Theorem 16).

Remind that the local maximum points $\mathbf{W} \in \mathcal{W}^{P \times K}$ are called *non-mixing* because they correspond to non-mixing transfer matrices from S to Y and thus to the recovering of P distinct sources. By contrast, the non-existence of *mixing* maxima (i.e. the maximum points satisfying $\mathbf{W} \in \mathcal{M}^{P \times K} \setminus \mathcal{W}^{P \times K}$) remains to be proved. Such a property, addressed in Chapter 3, ensures the equivalence between the local maximization of $\mathcal{C}_R(\mathbf{B})$ and partial source separation.

2.3.6 Support versus Range

As indicated in the beginning of Section 2.3.4, some results similar to those presented in the above subsection have been derived independently in [Cruces and Duran, 2004]. In their paper, the authors use an information theoretic approach to prove that under the $\|\mathbf{w}\| = 1$ constraint, $\mu[\Omega(Y_i)]$ reaches its minimum value when $Y_i \propto S_j$. From this, one can conclude that $-\mu[\Omega(Y_i)]$ is a deflation contrast. Clearly, this extends to $\mathcal{C}_R(\mathbf{b}_i)$ since it is equivalent to apply $-\mu[\Omega(Y_i)]$ on sources with convex support. From this viewpoint, Cruces' approach seems to be more interesting, because more general. However, there is no a priori reason to prefer using the support than the range. On the contrary, i) estimating the support is generally speaking more difficult than estimating the range because support estimation requires the computation of the extreme values of the rv (as for the range) as well as the location of the possible holes inside the support (not needed for the range computation), and ii) the range-based contrast benefits from an interesting property (namely, the *discriminacy* property) not shared by the support, as it will be shown in Chapter 3. This discriminacy property of the range-based criterion (in the sense used in this work), which was not mentioned in [Cruces and Duran, 2004] but first appeared in [Vrins et al., 2005a], shall be proved by using a kind of proof similar to those used in the above subsection.

2.3.7 A tool for building a D-BSS contrast based on Huber

In Section 2.1, Theorem 7 (p. 48) gives a result for building an orthogonal partial BSS contrast from deflation BSS contrasts and Theorem 8 (p. 50) gives two results for building simultaneous and partial BSS contrast functions from superadditive functionals. However, it is possible to extend the last results to deflation contrasts. For proving that $-h\left(Y/\sqrt{\text{Var}[Y]}\right)$ is a contrast function for deflation, two properties of the entropy power $2^{h(\cdot)}$ have been used: the EPI (Theorem 6 p. 38) and the fact that $2^{h(\alpha Y)} = |\alpha|2^{h(Y)}$. Therefore, since $-h(Y/\sqrt{\text{Var}[Y]})$ is a contrast, we have that for any strictly decreasing function Ψ , $\Psi\left[h\left(Y/\sqrt{\text{Var}[Y]}\right)\right]$ is a contrast and we conclude the following:

Corollary 10 Let $Q(\cdot)$ be a positive-valued class II superadditive functional in the sense of Huber as stated in Theorem 8 (p. 50). Then, for any strictly decreasing real-valued function Ψ ,

$$f(\mathbf{b}) = \Psi \left[Q \left(\mathbf{b}\mathbf{X} / \sqrt{\text{Var}[\mathbf{b}\mathbf{X}]} \right) \right] \quad (2.54)$$

is a deflation contrast. In particular, taking $\Psi[\cdot] = -\log[\cdot]$,

$$f(\mathbf{b}) = 1/2 \log \text{Var}[\mathbf{b}\mathbf{X}] - \log Q(\mathbf{b}\mathbf{X}) . \quad (2.55)$$

The global maximum is attained when $\mathbf{w} \propto \mathbf{e}_k$, where $k \doteq \underset{i \in \{1, \dots, K\}}{\operatorname{argmin}} Q(\mathbf{S}_i)$, that is if $\mathbf{b}\mathbf{X} \propto \mathbf{S}_k$. Note that the $\mathbf{w} \in \mathcal{S}(K)$ constraint implies $\text{Var}[\mathbf{b}\mathbf{X}] = 1$ under the unit-variance assumption on the sources \mathcal{A}_7 .

The proof of the corollary is similar to the proof of Lemma 5 (p. 38).

Setting $Q(\cdot) = 2^h(\cdot)$ and $\Psi[x] = -\log[x]$, we find the deflation contrast $C_h(\mathbf{b})$ given in Eq. (2.20) (remind that (Huber 1) results from Eq. (1.74) and (Huber 2) results from the EPI theorem, Theorem 6 p. 38). Taking $Q(\cdot) = R(\cdot)$ and $\Psi[x] = -x$, this proves the deflation contrast property of the range functional given in Eq. (2.44), which was previously proved in Theorem 14 p. 62 (remind that (Huber 1) results from Eq. (2.35) and (Huber 2) results from the BMI theorem, Theorem 13 p. 60). By setting $\Psi[x] = -\log[x]$, we recover the deflation version of the simultaneous and partial range-based criteria given in Eq. (2.41) and Eq. (2.51), respectively.

2.4 RÉNYI'S ENTROPY CONTRAST

The use of a generalized form of Shannon's entropy, called Rényi's entropy, has been proposed in Information Theoretic Learning because of its computational advantage on Shannon's entropy for specific values of $r \neq 1$ [Haykin, 2000], especially in speech [Flandrin et al., 1994] and image processing [Sahoo et al., 1997] as well as clustering [Jenssen et al., 2003], feature extraction [Hild et al., 2006a]. In particular, it has been proposed to solve the BSS problem [Erdogmus et al., 2002a, Hild et al., 2001, 2006b]. The motivation for doing so comes from the fact that support measure and Shannon's entropy are two specific cases of Rényi's entropy (with $r = 0$ and $r = 1$, respectively), and that setting $r = 2$ may help to simplify some calculations when Parzen windowing is used for density estimation [Parzen, 1962]. Indeed, if the pdf $p(x)$ is approximated by a sum of N Gaussian kernels

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} , \quad (2.56)$$

and

$$\hat{p}(x) = \sum_{n=1}^N \frac{\phi\left(\frac{x-\mu_n}{\sigma_n}\right)}{\sigma_n} , \quad (2.57)$$

then

$$\begin{aligned}
h_2[p] &= -\log \int p^2(x)dx \\
&\approx -\log \int \hat{p}^2(x)dx \\
&= -\log \left(N^{-2} \sum_{i=1}^N \sum_{j=1}^N \frac{1}{\sigma_i \sigma_j} \int \phi\left(\frac{x-\mu_i}{\sigma_i}\right) \phi\left(\frac{x-\mu_j}{\sigma_j}\right) dx \right) \\
&= 2 \log N - \log \left(\sum_{i=1}^N \sum_{j=1}^N \frac{1}{\sqrt{\sigma_i^2 + \sigma_j^2}} \phi\left(\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right) \right); \quad (2.58)
\end{aligned}$$

the integration vanishes because of the properties of Gaussian functions (see Section 2.6.8, p. 92). The main problem of this approach is that theoretical proofs ensuring that the sources will be recovered through the maximization of a criterion related to h_r , $r \notin \{0, 1\}$, are lacking; the justification of considering the general Rényi entropy as a BSS criterion is only based on simulation results. By using this general Rényi's entropy in BSS, several authors have implicitly conjectured that this quantity (with any $r > 0$) is a contrast function.

However, the use of the functionals $-h_0$ and $-h_1$ yields to contrast functions, because of the class II superadditive property of these functions (proved using the BMI and EPI, respectively). Therefore, it was first our hope to find a generalized form of these inequalities that would ensure that $e^{h_r(\cdot)}$ and/or $e^{h_{r,\Omega}(\cdot)}$ with arbitrary $r > 0$ would be a class II superadditive functional; if it was the case, one could take $Q(\cdot) = e^{h_r(\cdot)}$ or $Q(\cdot) = e^{h_{r,\Omega}(\cdot)}$. Unfortunately, we were not able to find such a unifying theorem. Then, a more neutral point of view had to be adopted: it is not a priori hoped that Rényi's entropy based criteria, generally speaking, can benefit from the contrast property. Instead of trying to prove that Rényi's entropy is a contrast, we shall check if some necessary conditions can be violated, preventing Rényi's entropy criteria to be contrast functions. In the deflation case, the criterion evaluated at a point \mathbf{b}^* corresponding to the extraction of the source with the lowest index value $Q(\cdot)$ (i.e. $\mathbf{b}^* \propto \mathbf{S}_j$ where $j = \operatorname{argmin}_{i \in \{1, \dots, K\}} Q(\mathbf{S}_i)$) must have a local maximum. Equivalently, for the simultaneous approach, the related criteria must face a global (i.e. at least a local) maximum at any point $\mathbf{W} \in \mathcal{W}(K)$. This study is handled in the next section via a Taylor development of the criteria and analyzing the two first-order terms. Note that as the $r = 0$ case has been studied separately, we can focus on $r \in \mathbb{R}_+^0$, implying $h_r = h_{r,\Omega} = h_{r,\bar{\Omega}}$.

2.4.1 Taylor development of Rényi's entropy

In this section, we adopt a similar approach as in [Pham, 2005] to extend the expansion of Shannon's entropy to the expansion of Rényi's entropy, which is obviously supposed to be finite.

Let Z be a random variable, possibly depending on Y , and ϵ be a small scalar. From the definition of Rényi's entropy given in Eq. (1.106), it comes that Rényi's entropy of $Y + \epsilon Z$ is

$$h_r(Y + \epsilon Z) = \frac{1}{1-r} \log \int p_{Y+\epsilon Z}^r(\xi) d\xi , \quad (2.59)$$

where the density $p_{Y+\epsilon Z}$ reduces to, up to first order in ϵ [Pham, 2005]:

$$p_{Y+\epsilon Z}(\zeta) = p_Y(\zeta) - \epsilon(E[Z|Y=y]p_Y(y))'|_{y=\zeta} + o(\epsilon) . \quad (2.60)$$

In the above equation, we have used the “small $o()$ ” Landau notation, where the argument is implicitly supposed to tend to zero: we say that $a(x) = o(x)$ if $\lim_{x \rightarrow 0} a(x)/x \rightarrow 0$ (i.e. $a(x)$ tends faster to zero than x). Similarly, the “big O ” Landau notation will be used in this work: $a(x) = O(x)$ means $|\lim_{x \rightarrow 0} a(x)/x| < \infty$ (as an example, $3x^2 = o(x)$ and $2x = O(x)$).

In Eq. (2.60), $(E[Z|Y=y]p_Y(y))'|_{y=\zeta}$ stands for the derivative of $E[Z|Y=y]p_Y(y)$ with respect to y evaluated at $y = \zeta$. Hence, noting that:

$$\begin{cases} \log(1 + \epsilon) = \epsilon + o(\epsilon) , \\ p_{Y+\epsilon Z}^r(\zeta) = p_Y^r(\zeta) - r\epsilon p_Y^{r-1}(\zeta)(E[Z|Y=y]p_Y(y))'|_{y=\zeta} + o(\epsilon) , \end{cases}$$

equations (2.59) and (2.60) yield

$$\begin{aligned} h_r(Y + \epsilon Z) &= \frac{1}{1-r} \log \left\{ \int p_Y^r(\xi) d\xi \right. \\ &\quad \left. - \int r\epsilon p_Y^{r-1}(\xi)(E[Z|Y=y]p_Y(y))'|_{y=\xi} d\xi \right\} + o(\epsilon) \\ &= \frac{1}{1-r} \log \int p_Y^r(\xi) d\xi \\ &\quad + \frac{1}{1-r} \log \left\{ 1 - \frac{\epsilon r \int p_Y^{r-1}(\xi)(E[Z|Y=y]p_Y(y))'|_{y=\xi} d\xi}{\int p_Y^r(\xi) d\xi} \right\} + o(\epsilon) \\ &= h_r(Y) - \frac{\epsilon r}{1-r} \frac{\int p_Y^{r-1}(\xi)(E[Z|Y=y]p_Y(y))'|_{y=\xi} d\xi}{\int p_Y^r(\xi) d\xi} + o(\epsilon) \quad (2.61) \end{aligned}$$

where we have used $\log(1 + a\epsilon + o(\epsilon)) = \log(1 + a\epsilon) + \log(1 + o(\epsilon)) = a\epsilon + o(\epsilon)$ when $\epsilon \rightarrow 0$. Note that this chain of equality requires that one can exchange limit and integration, see Rem. 13.

By integration by parts, one gets for well-behaved densities

$$\frac{1}{r-1} \int p_Y^{r-1}(\zeta)(E[Z|Y=y]p_Y(y))'|_{y=\zeta} d\zeta = - \int p_Y^{r-1}(y) E[Z|Y=\zeta] p_Y'(\zeta) dy , \quad (2.62)$$

yielding

$$-\epsilon \frac{r}{1-r} \frac{\int p_Y^{r-1}(\xi)[E[Z|Y]p_Y(Y)]'(\xi) d\xi}{\int p_Y^r(\xi) d\xi} = -\epsilon r \frac{\int p_Y^{r-1}(\xi) E[Z|Y=\xi] p_Y'(\xi) d\xi}{\int p_Y^r(\xi) d\xi} . \quad (2.63)$$

From the general iterated expectation lemma (p. 208 of [Gray and Davisson, 2004]), the right-hand side of the above equality equals

$$-\epsilon r \frac{E[p_Y^{r-2}(Y)p'_Y(Y)Z]}{\int p_Y^r(y)dy} = \epsilon E[\psi_{Y,r}(Y)Z] , \quad (2.64)$$

if we define the r -score function $\psi_{Y,r}(Y)$ of Y .

$$\psi_{Y,r}(y) \doteq -\frac{rp_Y^{r-2}(y)p'_Y(y)}{\int p_Y^r(y)dy} = -\frac{1}{p_Y(y)} \frac{(p_Y^r)'(y)}{\int p_Y^r(y)dy} . \quad (2.65)$$

Using the last equality, we observe that the r -score shares two major properties (see Property 4 p. 57) of the 1-score defined in Eq. (2.28); namely:

$$\begin{cases} E[\psi_{Y,r}(Y)] &= 0 , \\ E[\psi_{Y,r}(Y)Y] &= 1 . \end{cases}$$

We have thus a first-order expansion of Rényi's entropy, expressed as a function of the r -score:

$$h_r(Y + \epsilon Z) = h_r(Y) + \epsilon E[\psi_{Y,r}(Y)Z] + o(\epsilon) . \quad (2.66)$$

We now perform a second-order expansion of h_r . To this end, consider the second-order expansion of $p_{Y+\epsilon Z}$ provided in [Pham, 2005] (Z is temporarily assumed to be zero-mean, but definitely supposed to be independent from Y in order to make the development easier):

$$p_{Y+\epsilon Z}(\zeta) = p_Y(\zeta) + \frac{1}{2}\epsilon^2 E[Z^2]p_Y''(\zeta) + o(\epsilon^2) , \quad (2.67)$$

and

$$p_{Y+\epsilon Z}^r(\zeta) = p_Y^r(\zeta) + \frac{1}{2}rp_Y^{r-1}(\zeta)\epsilon^2 E[Z^2]p_Y''(\zeta) + o(\epsilon^2) . \quad (2.68)$$

Therefore, since Rényi's entropy is not sensitive to translation we have, for $r > 0$:

$$\begin{aligned} h_r(Y + \epsilon Z) &= \frac{1}{1-r} \left[\log \int p_Y^r(y)dy \right. \\ &\quad \left. + \log \left\{ 1 + \frac{1/2r\epsilon^2 \int p_Y^{r-1}(y)E[Z^2]p_Y''(y)dy}{\int p_Y^r(y)dy} \right\} \right] + o(\epsilon^2) \\ &= h_r(Y) + \frac{\epsilon^2}{2} \underbrace{\frac{\int p_Y^{r-1}(y)p_Y''(y)dy}{\int p_Y^r(y)dy}}_{\doteq J_r(Y)} \text{Var}[Z] + o(\epsilon^2) , \end{aligned} \quad (2.69)$$

where $J_r(Y)$ is called the r -th order information of Y (see Rem. 13). By integration by parts, we have that

$$J_r(Y) = r \frac{\int p_Y^{r-2}(y)(p'_Y(y))^2 dy}{\int p_Y^r(y)dy} , \quad (2.70)$$

which is a positive quantity whatever is $r > 0$. Observe that the first order information reduces to $J_1(Y) = E[\psi_Y(Y)^2]$, which is precisely Fisher's information [Cover and Thomas, 1991].

Remark 13 (On approximating integral of functions) *Equation (2.61) says that there exists a function $\phi(\epsilon, \zeta)$ such that*

$$p_{Y+\epsilon Z}^r(\zeta) = p_Y^r(\zeta) - r\epsilon p_Y^{r-1}(\zeta)(E[Z|Y=y]p_Y(y))'|_{y=\zeta} + \phi(\epsilon, \zeta) ,$$

where $\lim_{\epsilon \rightarrow 0} \phi(\epsilon, \zeta)/\epsilon = 0$ (because $\phi(\epsilon, \zeta) = o(\epsilon)$). By integrating both sides of the above equation, we find that $\int p_{Y+\epsilon Z}^r(\zeta)d\zeta$ equals

$$\int p_Y^r(\zeta)d\zeta - r\epsilon \int p_Y^{r-1}(\zeta)(E[Z|Y=y]p_Y(y))'|_{y=\zeta}d\zeta + \int \phi(\epsilon, \zeta)d\zeta .$$

Therefore, we have implicitly conjectured in Eq. (2.61) that $\int \phi(\epsilon, \zeta)d\zeta = o(\epsilon)$. However, this is not true whatever $\phi(\epsilon, \zeta)$ is. The possible problem is that we have no guarantee that $\phi(\epsilon, \zeta)/\epsilon$ converges uniformly to zero; the convergence could be only pointwise. Formally, in order to prove that $\lim_{\epsilon \rightarrow 0} \int \phi(\epsilon, \zeta)d\zeta/\epsilon = 0$ knowing $\phi(\epsilon, \zeta) = o(\epsilon)$, it suffices, by the Lebesgue Dominated Convergence Theorem, that there exist $\epsilon^* > 0$ and an integrable function $\delta(\zeta) > 0$, such that for all $\zeta \in \mathbb{R}$ and all $|\epsilon| < \epsilon^*$, $|\phi(\epsilon, \zeta)/\epsilon| < \delta(\zeta)$.

In Eq. (2.61), this additional requirement is actually implicitly assumed to be fulfilled, but this might require conditions on the pdf of the random variables Y and Z , that are not detailed here (same applies to second-order considerations). However, when the expanded function is a “well-behaved density”, we conjecture that this should be true in most of cases; in particular, observe that the results found via theoretical considerations are confirmed by specific numerical experiments.

Note that the permutation between the integral and limit signs corresponds to the condition under which, practically, approximating the integral of a function can be done via integration of an approximated form of the function. Actually, this is what people do when they compute entropies via density estimation; generally, one guesses $h_r(Y) \approx \frac{1}{1-r} \log [\int \hat{p}_Y^r(y)dy]$, i.e. that $h_r[\hat{p}_Y]$ can be rend as close as possible to $h_r[p_Y]$ provided that $\hat{p}_Y^r(y)dy$ is sufficiently close to $p_Y^r(y)dy$.

2.4.2 Deflation approach

In this subsection, we consider the scale invariant criterion

$$C_{h_r}(\mathbf{b}) \doteq -h_r \left(\frac{\mathbf{b}\mathbf{X}}{\sqrt{\text{Var}[\mathbf{b}\mathbf{X}]}} \right) . \quad (2.71)$$

Clearly, maximizing the above quantity with respect to \mathbf{b} is equivalent to maximizing

$$\tilde{C}_{h_r}(\mathbf{w}) \doteq -h_r \left(\frac{\mathbf{w}\mathbf{S}}{\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]}} \right) \quad (2.72)$$

if $\mathbf{w} = \mathbf{b}\mathbf{A}$.

Note that in both criteria, the denominator can be omitted under the $\mathbf{w} \in \mathcal{S}(K)$ constraint since $\mathcal{C}_{h_r}(\mathbf{b}) = \tilde{\mathcal{C}}_{h_r}(\mathbf{b}\mathbf{A}) = -h_r\left(\frac{\mathbf{w}}{\|\mathbf{w}\|}\mathbf{S}\right)$.

Based on the Taylor expansion of Rényi's entropy, which is proved to have a null first-order term when \mathbf{Y} and \mathbf{Z} are independent (see Eq. (2.64)), we find the following result, proved in the Appendix of the chapter (Section 2.6.9, p. 93).

Lemma 7 (Basis vectors are stationary points of \mathcal{C}_{h_r}) *The criterion $\tilde{\mathcal{C}}_{h_r}$ admits a stationary point when $\mathbf{w} \in \{\pm\mathbf{e}_1, \dots, \pm\mathbf{e}_K\}$.*

A sufficient argument for proving that \mathcal{C}_{h_r} is not a deflation contrast function is to prove that one of the stationary points of Lemma 7 is a local minimum. Indeed, if this occurs, the associated source will never be extracted through its maximization.

Actually, a necessary condition for the function $\tilde{\mathcal{C}}_{h_r}(\mathbf{w})$ over the set $\mathcal{S}(K)$ to admit a local maximum at $\pm\mathbf{e}_j$ is that $J_r(\mathbf{S}_j) \geq 1$ and a sufficient condition is that this inequality is strict. More generally, one can write these conditions as $J_r(\mathbf{S}_j)\text{Var}[\mathbf{S}_j] \geq 1$ and $J_r(\mathbf{S}_j)\text{Var}[\mathbf{S}_j] > 1$, which are then independent from the source variances. To see that $J_r(\mathbf{S})\text{Var}[\mathbf{S}]$ is invariant to the scale of \mathbf{S} , it suffices to note $\mathbf{S} = \sigma_{\mathbf{S}}\mathbf{S}^*$ where $\sigma_{\mathbf{S}}^2 = \text{Var}[\mathbf{S}]$ and \mathbf{S}^* is the unit-variance copy of \mathbf{S} . Then, using the density of a transformation given in Eq. (1.76) and from the definition of J_r in Eq. (2.69), we find

$$J_r(\mathbf{S}) = \frac{1}{\sigma_{\mathbf{S}}^2} J_r(\mathbf{S}^*) . \quad (2.73)$$

Observe that in the specific $r = 1$ case, the sufficient condition $J(\mathbf{S}_j)\text{Var}[\mathbf{S}_j] > 1$ is always satisfied for non-Gaussian sources (see Property 4, p. 57).

From the second order development of Rényi's entropy, one gets the following lemma (see the proof in Section 2.6.10, p. 94).

Lemma 8 (Contrast condition for \mathcal{C}_{h_r} , deflation approach) *The criterion $\mathcal{C}_{h_r}(\mathbf{b})$ under the constraint $\text{Var}[\mathbf{b}_i\mathbf{X}] = 1$ is not a contrast if $J_r(\mathbf{S}_i)\text{Var}[\mathbf{S}_i] < 1$ where*

$$i \in \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} -h_r\left(\frac{\mathbf{S}_k}{\sqrt{\text{Var}[\mathbf{S}_k]}}\right) . \quad (2.74)$$

2.4.3 Simultaneous approach

The simultaneous criterion associated to $h_r(\cdot)$ is

$$\mathcal{C}_{h_r}(\mathbf{B}) \doteq \log |\det \mathbf{B}| - \sum_{i=1}^K h_r(\mathbf{b}_i\mathbf{X}) . \quad (2.75)$$

The criterion is subject to the normalization constraint that $\mathbf{b}_i \Sigma_{\mathbf{X}} \mathbf{b}_i^T = 1$ (this is only to enforce the recovering of unit-variance outputs, the above criterion is not sensitive to the scale of \mathbf{b}_i).

We have the following result (see Section 2.6.11, p. 95 in the Appendix of the chapter):

Lemma 9 (Stationary points of \mathcal{C}_{h_r} , simultaneous approach) *The criterion \mathcal{C}_{h_r} admits a stationary point when $\mathbf{BA} \in \mathcal{W}(K)$ or, equivalently, when $\mathbf{B} \sim \mathbf{A}^{-1}$.*

We now derive in the next lemma (proved in Section 2.6.12, p. 95) a necessary and sufficient condition for the criterion $\mathcal{C}_{h_r}(\mathbf{B})$ to attain a local maximum at the point $\mathbf{B} \sim \mathbf{A}^{-1}$ (and consequently, a sufficient condition ensuring that $\mathcal{C}_{h_r}(\mathbf{B})$ is not a contrast function).

Lemma 10 (Contrast condition for \mathcal{C}_{h_r} , simultaneous approach) *The criterion $\mathcal{C}_{h_r}(\mathbf{B})$ is not a contrast if the sources share a same density p_S and $J_r(S)\text{Var}[S] < 1$ where S is a random variable with density p_S .*

2.4.4 Partial approach

As the simultaneous and deflation approaches are particular cases of the partial separation, the results presented in the above subsections show that, generally speaking, ERE is not a contrast function for BSS.

2.4.5 Numerical simulation and detailed calculation on specific examples

Lemma 8 p. 71 and Lemma 10 p. 72 give sufficient conditions ensuring that the above deflation and simultaneous criteria are *not* contrast functions: maximizing them will not lead to recover the sources if these conditions are met. It is shown in this section that these conditions can easily be encountered for densities close to (but different from) Gaussian functions and for specific values of Rényi's exponent r .

Consider the case where the common density of the source admits a density of the form

$$p_S(s) = C e(-|s/\lambda|^a/a) , \quad (2.76)$$

where a is a positive parameter, λ is a positive scale parameter and C is the normalizing constant. Then, S denoting a random variable with density p_S and denoting the r -score function of S evaluated at the point y by $\psi_{S,r}(y)$, simple manipulations yield

$$\psi_{S,r}(y) = \frac{r \operatorname{sign}(y)|y|^{a-1}\lambda^{-a}}{Ce[-(1-r)|y/\lambda|^a/a]\int e(-r|u/\lambda|^a/a)du}, \quad (r > 0). \quad (2.77)$$

In particular, $\psi_{S,1}(y) = \text{sign}(y)|y|^{a-1}\lambda^{-a}$. Further

$$J_r(S) = \frac{r \int |s|^{2a-2} \lambda^{-2a} e(-r|s/\lambda|^a/a) ds}{\int e(-r|s/\lambda|^a/a) ds} \quad (2.78)$$

$$= \frac{r \int |u|^{2a-2} e(-r|u|^a/a) du}{\lambda^2 \int e(-r|u|^a/a) du} \quad (2.79)$$

$$= \frac{r^{2/a-1} \int |z|^{2a-2} e(-|z|^a/a) du}{\lambda^2 \int e(-|z|^a/a) dz} = \frac{r^{2/a-1}}{\lambda^2} E|Z|^{2a-2} \quad (2.80)$$

where $Z = S/\lambda$ is a random variable with density $e(-|z|^a/a)/\int e(-|u|^a/a) du$. Since $\text{Var}[S] = \lambda^2 E[Z^2]$, one has

$$J_r(S)\text{Var}[S] = r^{2/a-1} \underbrace{E[|Z|^{2a-2}]E[Z^2]}_{\doteq g(a)}, \quad (r > 0), \quad (2.81)$$

which is independent from the scale parameter λ as it should be. In particular, for $a = 2$, which corresponds to S and Z being Gaussian with $E[Z^2] = 1$, one has $J_r(S)\text{Var}[S] = 1, \forall r > 0$.

Put $g(a) = E[|Z|^{2a-2}]E[Z^2]$, which from the above result equals $J(S)\text{Var}[S]$ where $J(S) = J_1(S)$ is no other than Fisher's information of S . But we know that $J(S)\text{Var}[S] \geq 1$ with equality if and only if S is Gaussian, that is $a = 2$. Thus g admits a global minimum equal to 1 at $a = 2$. Explicitly from the definition of the gamma function: $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$, one has

$$\begin{aligned} E[|Z|^\beta] &= \frac{\int_0^\infty e(-z^a/a) z^\beta dz}{\int_0^\infty e(-z^a/a) dz} \\ &= \frac{\int_0^\infty e(-t)(at)^{(\beta+1)/a-1} dt}{\int_0^\infty e(-t)(at)^{1/a-1} dt} \\ &= a^{\beta/a} \frac{\Gamma[(\beta+1)/a]}{\Gamma(1/a)}. \end{aligned} \quad (2.82)$$

Therefore, defining

$$g(a) = E[|Z|^{2a-2}]E[Z^2] = a^2 \frac{\Gamma(2-1/a)\Gamma(3/a)}{\Gamma(1/a)^2}, \quad (2.83)$$

one can check that g admits indeed a global minimum at $a = 2$.

Finally $J_r(S)\text{Var}[S] < 1$ if and only if

$$r < g(a)^{1/(1-2/a)} < 1 \quad \text{in the case } a < 2, \quad (2.84)$$

$$r > g(a)^{1/(1-2/a)} > 1 \quad \text{in the case } a > 2. \quad (2.85)$$

One concludes that for source densities of the form $p_S(s) = Ce(-|s/\lambda|^a/a)$, if $a < 2$ then the criteria are not contrasts for $r < g(a)^{a/(a-2)}$ and if $a > 2$, they

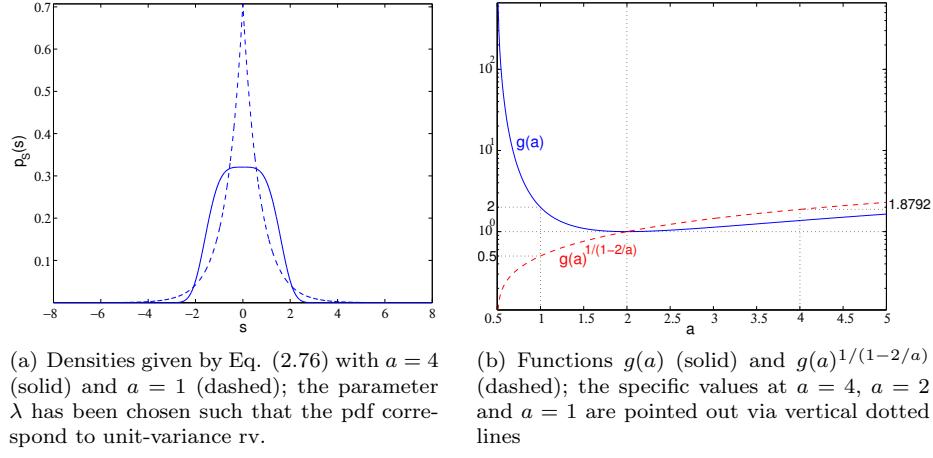


Figure 2.2. Discussion example (see text).

are not contrasts for $r > g(a)^{a/(a-2)}$. In particular, for bilateral exponential sources, which correspond to $a = 1$, one has $g(a) = 2$ and thus the criteria are not contrasts for $r < 1/2$. For $a = 4$, $g(a)^{1/(1-2a)} = 1.8792$ and thus the criteria are not contrasts for $r > 1.8792$, in particular for $r = 2$. The densities (2.76) with $a = 1$ and $a = 4$ as well as the functions $g(a)$ and $g(a)^{1/(1-2a)}$ are illustrated in figures 2.2.(a) and 2.2.(b)

The approximated Rényi entropy $\bar{h}_r(Y_\theta)$ (see Rem. 14 below) where $Y_\theta = \mathbf{w}_\theta S$ as a function of the transfer angle θ for the two above examples is illustrated in figures 2.3.(a) and 2.3.(b). The two unit-variance sources share the same density : $p_{S_1} = p_{S_2} = p_S$ where p_S is given by Eq. (2.76). Figure 2.4. shows the case where the source shape parameters are different: $a_{S_1} = 4$ and $a_{S_2} = 1$.

Remark 14 (Some details regarding the simulation method) The estimation $\bar{h}_r(Y_\theta)$ of $h_r(Y_\theta)$ is defined as

$$\bar{h}_r(Y_\theta) = \begin{cases} \frac{1}{1-r} \log \sum_{\Delta} [p_{\sin \theta S} * p_{\cos \theta S}]^r & \text{if } \theta \notin \{k\pi/2, k \in \mathbb{Z}\} \\ \frac{1}{1-r} \log \sum_{\Delta} p_{S_i}^r & \text{otherwise (} i^* \text{ depends on } \theta \text{),} \end{cases}$$

where the \sum_{Δ} symbol denotes the Riemannian approximation of the exact integral (the step Δ is taken equal to 10^{-3} and the grid size is chosen large enough to ensure that the integration error is limited, $\max(|1 - \sum_{\Delta} p_{\sin \theta S}|, |1 - \sum_{\Delta} p_{\cos \theta S}|) < \tau$, $\tau = 1E^{-4}$ and similarly, the variance deviation error is also controlled $\max(|1 - (\sum_{\Delta} s^2 p_{\sin \theta S}(s))|, |1 - \sum_{\Delta} s^2 p_{\cos \theta S}(s)|) < \tau$). The exact theoretical expressions of $p_{\sin \theta S}$ and $p_{\cos \theta S}$ have been dealt with and the convolution operation is performed via the Matlab `conv` command. When computing (2.86), the summation inside the log is only computed on discrete points s_0 satisfying

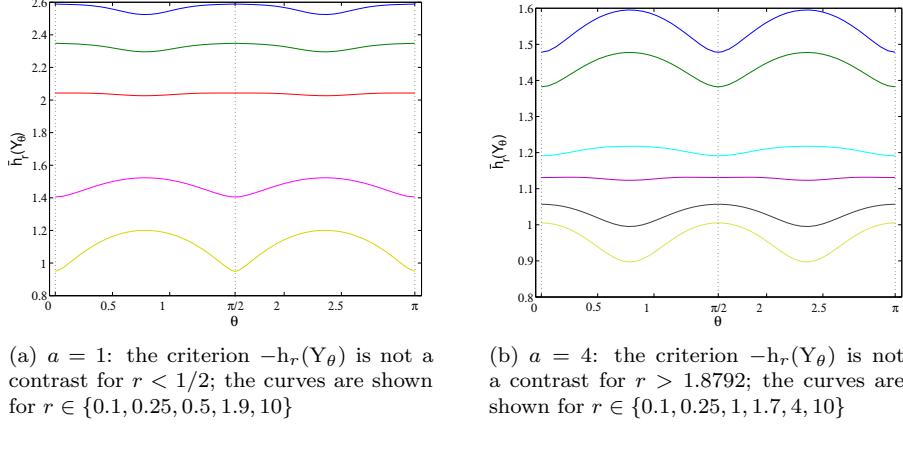


Figure 2.3. Evolution of $\bar{h}_r(Y_\theta)$ where $ps_1 = ps_2 = ps$ is given by Eq. (2.76) with $\lambda = 1$; (remind that h_r is decreasing in r).

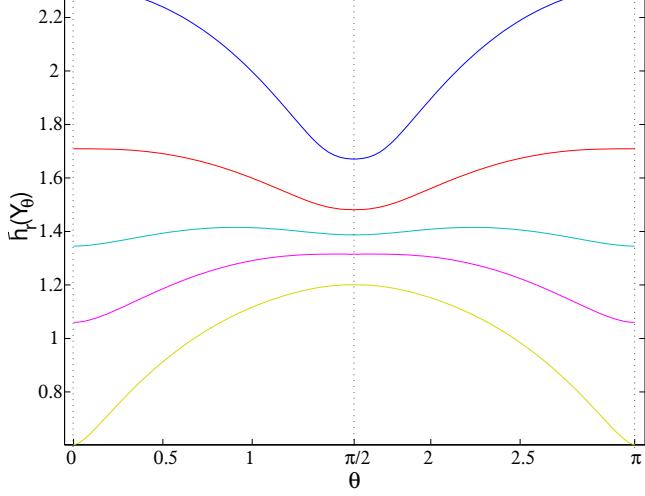


Figure 2.4. Evolution of $\bar{h}_r(Y_\theta)$ where ps_1 by Eq. (2.76) with $a = 4$ and ps_2 by Eq. (2.76) with $a = 1$. The kind of non-mixing optimum of the criterion $-\bar{h}_r(Y_\theta)$ depends on the source that is extracted; the curves are shown for $r \in \{0.1, 0.5, 1, 1.9, 10\}$ (remind that h_r is decreasing in r).

$p_{Y_\theta}(s_0) > \tau$ in order to avoid to face numerical problems resulting from a value close to $\log(0)$.

Remark 15 Consider the function $a \mapsto r^{2/a-1}g(a) = J_r(S)\text{Var}[S]$. It takes the value 1 at 2 and its logarithmic derivative is $-2a^{-2}\log r + g'(a)/g(a)$, which

takes the value $-\frac{1}{2} \log r$ at $a = 2$ (since g is minimum at 2). Thus, for $r < 1$, this function is increasing in a neighborhood of 2, hence there exists an $a < 2$ for which $J_r(S)\text{Var}[S] < 1$. Similarly, for $r > 1$, this function is decreasing in the neighborhood of 2, hence there exists an $a > 2$ for which $J_r(S)\text{Var}[S] < 1$. Thus for any $r \neq 1, r > 0$, there exists a sources density of the form $Ce(-|s/\lambda|^a/a)$ for some a for which the criterion is not a contrast. As Rényi's entropy power is a class II functional and because this condition combined with its possible strict superadditivity necessarily implies that \mathcal{C}_{h_r} is a contrast function (from Theorem 8, p. 50), we conclude the following:

Corollary 11 For any $r > 0, r \neq 1$, there always exists a pair of i.i.d. random variables with common density belonging to the generalized exponential family (but differing from the Gaussian function) such that the r -Rényi entropy power cannot be a superadditive functional for these variables.

Remark 16 Figures 2.3. and 2.4. seem to indicate that even if the kind of the extremum points changes with r (maximum or minimum), their location is constant with respect to Rényi's exponent. This is not the case, generally speaking. Let us focus on the deflation criterion with $K = 2$ and restrict ourselves to $\theta \in [0, \pi/2]$. It has been shown that a stationary point always exists when $\theta = \pi/2$, whatever r . Simple calculations yield:

$$Y_{\theta+\delta\theta} = Y_\theta + [\delta\theta \cos \theta, -\delta\theta \sin \theta] S + o(\delta\theta) . \quad (2.86)$$

From Eq. (2.132), this leads to

$$h_r(Y_{\theta+\delta\theta}) = h_r(Y_\theta) - \delta\theta (\cos \theta E[\psi_{Y_\theta, r}(Y_\theta) S_1] - \sin \theta E[\psi_{Y_\theta, r}(Y_\theta) S_2]) + o(\delta\theta) . \quad (2.87)$$

Consequently, $h_r(Y_\theta)$ admits a stationary point at θ^* if

$$\tan \theta^* = \frac{E[\psi_{Y_\theta, r}(Y_\theta) S_1]}{E[\psi_{Y_\theta, r}(Y_\theta) S_2]} + o(\delta\theta) \quad (2.88)$$

$$= \frac{\int (p_{Y_\theta}^r)'(y) E[S_1 | Y_\theta = y] dy}{\int (p_{Y_\theta}^r)'(y) E[S_2 | Y_\theta = y] dy} + o(\delta\theta) . \quad (2.89)$$

The value of θ^* , generally speaking, depends on r . But if $p_{S_1 | Y_{\theta^*}} = p_{S_2 | Y_{\theta^*}}$, the above conditional expectations are identical and the ratio in the right-hand part of the above equation is always equal to one. This indicates that a stationary point exists at $\theta^* = \pi/4$ if $p_{S_1} = p_{S_2}$, whatever is $r > 0$.

As a last example, consider the case where the common density of the sources has the triangular density $p_T(s) = 1 - |s|$ if $|s| \leq 1, = 0$ otherwise. Then, denoting by S a random variable with the triangular density p_T , we have

$$\text{Var}[S] = 2 \int_0^2 (1-s)s^2 ds = \frac{1}{6} , \quad (2.90)$$

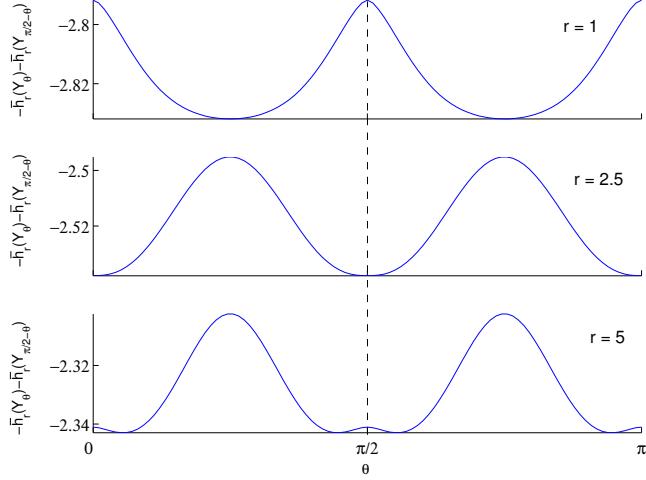


Figure 2.5. Evolution of estimated Rényi's criterion $-\bar{h}_r(Y_\theta) - \bar{h}_r(Y_{\pi/2-\theta})$ as a function of the transfer angle θ where the two sources share the same triangular density p_T . The criterion with $r = 2.5$ and $r = 5$ is not a contrast function.

and

$$J_r(S) = r \frac{\int_0^1 (1-s)^{r-2} ds}{\int_0^1 (1-s)^r ds} = r \frac{\int_0^1 u^{r-2} du}{\int_0^1 u^r du} = \begin{cases} r(r+1)/(r-1) & \text{if } r > 1 \\ \infty & \text{if } r \leq 1 \end{cases} \quad (2.91)$$

Thus $J_r(S)\text{Var}[S] < 1$ if and only if $r > 1$ and $r(r+1)/[6(r-1)] < 1$. But for $r \geq 1$, the last inequality is equivalent to $0 > r(r+1) - 6(r-1) = (r-2)(r-3)$. Therefore $J_r(S)\text{Var}[S] < 1$ if and only if $2 < r < 3$. We conclude that for triangular source, the criteria are not contrast functions if $2 < r < 3$.

Regarding the simultaneous criterion, the last two plots of Figure 2.5. clearly indicate that the problem could be emphasized too. On top of the figure ($r = 1$), the criterion $C_{h_r}(\mathbf{B}) = C_h(\mathbf{B})$ is a contrast function, as expected. On the middle plot ($r = 2.5$), $\tilde{C}_{h_r}(\mathbf{B})$ admits a local minimum point when $\theta \in \{k\pi/2 : k \in \mathbb{Z}\}$ (this results from $J_r(S)\text{Var}[S] < 1$), and thus violates a necessary requirement for a contrast function. Finally, on the last plot ($r = 5$), the criterion is not a contrast even though $J_r(S)\text{Var}[S] > 1$ since the set of *global* maximum points of the criterion does not correspond to the set $\mathcal{W}(K)$.

Remark 17 (On the nature of Hartley's and Shannon's entropies)

Some arguments have been given in the literature to emphasize the specific properties of Shannon and/or Hartley's entropies in the class of the generalized entropies. Some are based on “question-assertion” considerations [Knuth, 2005], others on average information gain/loss [Rényi, 1976b] or yet on related inequalities [Costa and Cover, 1984]. Without entering the details, it is explained in [Aczel et al., 1974] that, only linear combinations of Shannon and Hartley's entropies correspond to a “natural behavior”. Unfortunately, all those explanations

rely on discrete processes only. Moreover, their connections with the aforementioned “complexity measure meaning” remains unclear. Up to now, there is no really convincing “philosophical” results explaining why Shannon and Hartley’s entropies differ from other ones in the BSS context.

2.5 CONCLUSION OF THE CHAPTER

2.5.1 Summary of results

It was suggested in Chapter 1 that just as “independence measures” can yield contrast functions, the generalized form of Rényi’s information measures can also be the genesis of new contrast functions. This was motivated by the suitable “complexity measure” behavior of Shannon’s entropy; it was proposed by several authors in the past, even if a detailed theoretical study was missing. Therefore, Chapter 2 aims at filling this lack and thus at analyzing the entropic criteria, and more explicitly, their global maximum points. Indeed, the main property of contrast functions concerns the location of these global maximum points. For the deflation and partial separation schemes, a more general study of the local maximum points corresponding to transfer vectors proportional to basis vectors or to transfer matrices subPD-equivalent to the identity matrix was managed.

Instead of showing that the global maximum of a criterion corresponds to a non-mixing point, some specific tools can be used, such as Pham’s theorem (Theorem 8 p. 50): it suffices that the criterion has a specific form and satisfies a superadditivity condition to ensure that the criterion is a contrast. The criteria based on Shannon’s entropy, on the support or on the range are shown to fulfill this criterion. Further, the superadditivity conditions directly result from well-known inequalities: the entropy power inequality (Shannon’s entropy) and the Brunn-Minkowski inequality (range and support). In a more general way however, we were not able to find an extension of the EPI and BMI suggesting the superadditivity of Rényi’s entropy powers whatever the value of Rényi’s exponent r . Afterwards, this is logical. Based on a Taylor expansion of Rényi’s entropy, a sufficient condition for Rényi’s entropy-based criteria not being contrast functions was found, and some counter-examples illustrate that this condition is met in simple situations. Whatever is $r \notin \{0, 1\}$, there always exist a non-Gaussian density ($a \neq 2$) of the generalized exponential family such that the r -Rényi entropy-based criterion is not a contrast function if the sources follow this density; this was stated in Rem. 15 (p. 76). Surprisingly, this gives a partial answer to the question about the possible superadditivity of Rényi’s entropy, that remained an open question up to now. These results are summarized in Table 2.1. The “KO” results are proved via theoretical counterexamples showing that the corresponding property might be violated in some cases (even under the usual non-Gaussianity assumption). The “–” superscript indicates that these results are unexpected (but not contradictory) compared to the literature.

	Deflation	Simultaneous	Partial
Shannon ($r = 1$)	OK	OK	OK
Hartley ($r = 0$)	OK	OK	OK
Rényi ($r > 0, r \neq 1$)	KO ⁻	KO ⁻	KO ⁻
Range (ext. Hartley)	OK	OK	OK

Table 2.1. Summary of the results of Chapter 2: analysis of the contrast property of entropy-based criteria for the deflation, simultaneous and partial BSS. It is rigorously proved that Shannon, Hartley and extended Hartley entropies all yield to contrast function for the three separation schemes. By contrast, it always exist counter-examples showing that Rényi's entropy might be not a contrast function whatever is $r > 0, r \neq 1$. Original results are boldfaced, alternative proofs have been used to prove the known results.

2.5.2 Use of Rényi entropies in blind separation/deconvolution

How do our conclusions match with existing results ? The general form of Rényi's entropies have been proposed for blind source separation in [Erdogmus et al., 2002a, Hild et al., 2001, 2006b, Principe et al., 2000]. Computational convenience and close relationship with Shannon's entropy were the principal motivations and justifications for their use. Our conclusions seem to be in complete contradiction with these results. Indeed, even if Rényi's exponent is set to its more convenient value $r = 2$ (corresponding to the so-called *quadratic entropy*, avoiding thus the integration of a Gaussian product if Parzen density estimation using Gaussian kernels is used), some counter-examples show that the associated criterion is not always a contrast function, depending on some specificities of the source densities. The apparent contradiction is very simple to explain: whereas the authors of the above-referenced papers based their conclusions on simulation results (involving thus specific source densities and Rényi's exponents), our approach deals with a more general theoretical development. Therefore, even if in practice specific r -Rényi's entropies can be used for BSS (depending on the case), it is not generally speaking, a good BSS criterion.

Based on arguments that technically sound better, Rényi's entropies were also proposed for blind deconvolution [Erdogmus et al., 2002b, 2004, Bercher and Vignat, 2002]. We sketch below one of the approaches justifying their use, due to Bercher and Vignat in 2002. The starting point is the strengthened Young's inequality [Barthe, 1998, Gardner, 2002]:

Theorem 17 (Strengthened Young's inequality) *Let $\min(p, q, r) > 0$ and $1/p + 1/q = 1 + 1/r$, and let $f \in L^p(\mathbb{R}^N)$, $g \in L^q(\mathbb{R}^N)$ be non-negative functions. Finally, define $C_t \doteq \sqrt{\frac{|t|^{1-t}}{|t'|^{1/t'}}$ where t' is the Hölder complement of t , that is*

$1/t + 1/t' = 1$. Then:

$$\text{if } \min(p, q, r) \geq 1 : \|f * g\|_r \leq \left(\frac{C_p C_q}{C_r} \right)^N \|f\|_p \|g\|_q , \quad (2.92)$$

$$\text{if } \max(p, q, r) \leq 1 : \|f * g\|_r \geq \left(\frac{C_p C_q}{C_r} \right)^N \|f\|_p \|g\|_q . \quad (2.93)$$

In this theorem, $\|f\|_p \doteq \sqrt[p]{\int f^p(x)dx}$ and “*” denotes the convolution product. According to [Gardner, 2002], the first inequality was independently proven in [Beckner, 1975] as well as in [Brascamp and Lieb, 1976], and the second one appeared in [Brascamp and Lieb, 1976].

Assume that f and g are the densities of $w_1 S_1$ and $w_2 S_2$, respectively, where the S_i are independent non-Gaussian random variables. Then, obviously, $h_r[f * g] = h_r(Y)$. Setting $p = r$, $q = 1$ and noting that $\|g\|_1 = 1$ because g is a density, we find $\log \|f\|_r = -1/r' h_r(w_1 S_1)$, $\log \|f * g\|_r = -1/r' h_r(Y)$, where r' is the Hölder complement of r [Dembo et al., 1991]. As this must remain true when f and g are exchanged, we find by using the monotonicity of the logarithm and the class II property of Rényi entropy power that, if the sources are i.i.d. with common Rényi's entropy noted $h_r(S)$:

$$h_r(Y) \geq h_r(S) + \log \max |w_i| . \quad (2.94)$$

This is the essence of the inequality (7) in [Bercher and Vignat, 2002]. This condition is very weak actually. For instance, this inequality is not strong enough to prove that Rényi's entropy leads to contrast functions under a fixed variance constraint on the output. Let us see that. Under a fixed variance constraint, we can note $\mathbf{w} = \mathbf{w}_\theta$, $Y_\theta = \mathbf{w}_\theta S$ and the above inequality becomes

$$h_r(Y_\theta) \geq \max(h_r(S) + \log |\sin(\theta)|, h_r(S) + \log |\cos(\theta)|) . \quad (2.95)$$

Now, have a look at Figure 2.6.(a), in which the curves $h_r(\sin(\theta)S_1) = h_r(S) + \log |\sin(\theta)|$ and $h_r(\cos(\theta)S_2) = h_r(S) + \log |\cos(\theta)|$ have been plotted as a function of the transfer angle θ . Inequality (2.94) states that these curves lower-bound $h_r(Y_\theta)$. This inequality does not imply, unfortunately, that the minimum value of $h_r(Y)$ is reached when $\theta \in \{k\pi/2 | k \in Z\}$, i.e. minimizing $h_r(Y_\theta)$ according to θ might not lead to source recovering. A simple counter example is provided on the figure; the shape of the $h_r(Y_\theta)$ curve shown on the figure does not violate neither the above inequality nor the strict equality condition at the boundaries of the quadrant. However, the global minimum of this curve does not lead to $Y_\theta \in \{\pm S_1, \pm S_2\}$! More precisely, provided that the source entropies are finite, we have $\max(h_r(\sin(\theta)S_1), h_r(\cos(\theta)S_2)) < h_r(S)$ for all $\theta \in]0, \pi/2[$.

In summary, on the one hand [Erdogmus et al., 2002b, 2004] and [Bercher and Vignat, 2002] claimed that (2.94) justifies the use of Rényi entropies for the deconvolution of stationary sources (and indirectly for the separation of i.i.d. sources from linear instantaneous mixtures) and, on the other hand, we prove

here that this inequality is not strong enough to validate the method, even under the i.i.d. assumption on the sources. So, who is wrong ? Actually, the apparent contradiction between these results comes from a confusion about the normalization constraint.

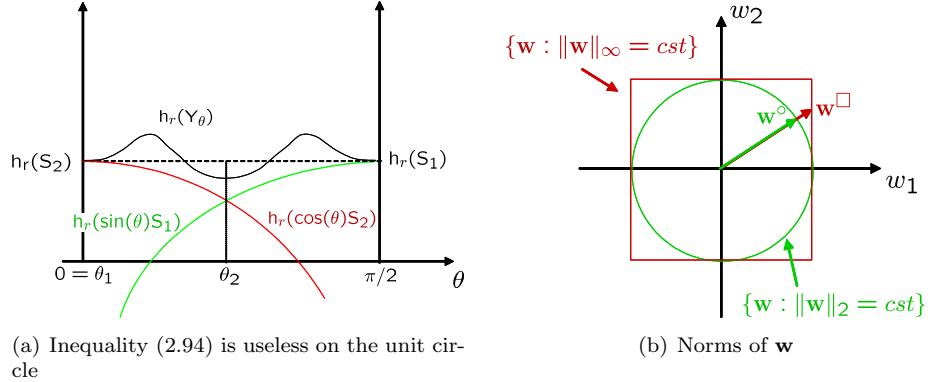
To show that inequality (2.94) is of no use, we have rather considered the inequality (2.95), which is nothing but (2.94) in which the $\|\mathbf{w}\| = 1$ constraint has been plugged. Throughout this chapter, the same constraint was used so that our conclusion is “Rényi’s entropy *under a $\|\mathbf{w}\| = 1$ constraint*, or similarly, *under a unit-variance constraint on the output* should not be used for BSS, even if the sources are i.i.d.”. In that sense, we disagree with the suggestion made in [Erdogmus et al., 2002b, 2004], where the authors justify the use of Rényi’s entropy based on inequality (2.94) combined to the fixed variance constraint (i.e. based on (2.95) if $K = 2$). Rigorously speaking however, our results are not in contradiction with the results of Bercher and Vignat, who proposed to use Rényi’s entropy but with a different normalization constraint.

Geometrically, our conclusion means that contrarily to the Shannon and (extended) Hartley cases, minimizing Rényi’s entropy with $r > 0$, $r \neq 1$ over the unit circle (in the space of $1 \times K$ transfer vectors) is not a good idea. But how does this conclusion extend to other search spaces ? For example, what if the constraint becomes $\|\mathbf{w}\|_p = cst$ with $p \neq 2$ (remind that by definition, $\|\mathbf{w}\| = \|\mathbf{w}\|_2$)? In [Bercher and Vignat, 2002], the authors proposed to set $p = \infty$. Let us illustrate graphically why this is *a priori* a good idea. To that aim, define $\mathbf{w}^\circ \in \{\mathbf{w} : \|\mathbf{w}\| = 1\}$ and $\mathbf{w}^\square \in \{\mathbf{w} : \|\mathbf{w}\|_\infty = 1\}$ where the superscript symbols refer to the geometry of the set of vectors with the associated p -norm (see Fig. 2.6.(b)). For comparison purpose, we assume that $\mathbf{w}^\circ = \mathbf{w}^\square / \|\mathbf{w}^\square\|$. It is clear that $\|\mathbf{w}^\square\| \geq 1$, and the equality case corresponds to vectors $\mathbf{w}^\circ = \mathbf{w}^\square = \pm \mathbf{e}_i$. Similarly, we define $\mathbf{Y}^\circ = \mathbf{Y}_\theta = \mathbf{w}^\circ \mathbf{S}$ and $\mathbf{Y}^\square = \mathbf{w}^\square \mathbf{S}$. Hence:

$$\begin{aligned} h_r(\mathbf{Y}^\circ) &= h_r(\mathbf{w}^\circ \mathbf{S}) \\ &= h_r(\mathbf{w}^\square \mathbf{S}) - \log \|\mathbf{w}^\square\| \\ &\leq h_r(\mathbf{Y}^\square) \end{aligned} \tag{2.96}$$

because $\|\mathbf{w}^\square\| \geq 1$. Actually, it is easily seen that $h_r(\mathbf{w} \mathbf{S})$ s.t. $\|\mathbf{w}\|_p = cst$ is always higher than $h_r(\mathbf{w} \mathbf{S})$ s.t. $\|\mathbf{w}\|_q = cst$ if $p \geq q$ (the strict equality is attained if and only if $\mathbf{w}^\circ = \pm \mathbf{e}_k$). Hence, the entropy of $h_r(\mathbf{w} \mathbf{S})$ s.t. $\|\mathbf{w}\|_p = cst$ increases with p for a given direction \mathbf{w} . Setting $p = \infty$ yields the $\|\mathbf{w}\|_p = cst$ search space on which $h_r(\mathbf{w} \mathbf{S})$ is maximum in a given direction. However, this function does not depend on the value of p at $\mathbf{w} = \pm \mathbf{e}_i$. Therefore, there is some chance that if some local maximum of $h_r(\mathbf{w} \mathbf{S})$ occurred on the unit circle at the basis vectors, they become local minimum points on search spaces of the form $\|\mathbf{w}\|_p$, $p > 2$, and the most favorable situation is obviously $p = \infty$.

By the EPI, we know that $h_1(\mathbf{Y}^\circ) \geq \min_i h_1(\mathbf{S}_i)$ with equality if and only if $\mathbf{w}^\circ = \pm \mathbf{e}_k$ with $k = \operatorname{argmin}_i h_1(\mathbf{S}_i)$. A similar conclusion can be drawn for $r = 0$ based on the BMI. The transitivity of the inequality and inequality (2.96)

**Figure 2.6.** Rényi's entropies and fixed p -norms search space.

yield $h_1(Y^\square) \geq \min_i h_1(S_i)$, $h_0(Y^\square) \geq \min_i h_0(S_i)$, with equality if and only if $\mathbf{w}^\square = \pm \mathbf{e}_k$. How does that extend to the general Rényi case ? By (2.94), we know that $h_r(Y^\square) \geq h_r(S)$ but, on the contrary, this is no more true when Y° is considered instead of Y^\square in the inequality, as shown in Section 2.4.5.

Figures 2.3.(a) and 2.3.(b) have been redrawn in a 3D space under the $\|\mathbf{w}\|_\infty = 1$ and $\|\mathbf{w}\| = 1$ constraints (Fig. 2.7.(a) and Fig. 2.7.(b)) for comparison purposes. One can see that the non-mixing local minima (located by ‘o’ markers) of $h_r(Y^\square)$ are strengthened compared to those of $h_r(Y^\circ)$ and that the non-mixing local maxima (located by ‘*’ markers) of $h_r(Y^\circ)$ vanish when Y^\square is considered instead of Y° . Even if this cannot be proved by using the above inequality (2.94), this extends to the example of Figure 2.4. (where $h_r(S_1) \neq h_r(S_2)$), as shown in Fig. 2.7.(c)

The question then becomes: how do our results extend from $\|\mathbf{w}\| = cst$ to $\|\mathbf{w}\|_\infty = cst$? The existing results provided in [Bercher and Vignat, 2002] partially answer this question when the sources are i.i.d, but what if they have very different Rényi entropies ? A preliminary question is the following: is it possible to perform the optimization over the $\|\mathbf{w}\|_p = cst$ constraint ? This is clearly possible for $p = 2$ by constraining the output variance to be constant; but is it possible do that for $p = \infty$ based on the elements of the demixing matrices only? The answer is negative. The theoretical results of Bercher & Vignat are correct, but there is no way to fulfill the $\|\mathbf{w}\|_p = cst$ in practice (i.e. knowing the demixing matrix and the mixtures only) if $p \neq 2$, making their developments useless. Therefore, our negative conclusions about the use of Rényi's entropy in BSS still hold.

The aim of the next chapter is to deal with the more difficult problem of spurious local maxima, i.e. local maxima at transfer vectors that are not proportional to any of the basis vectors.

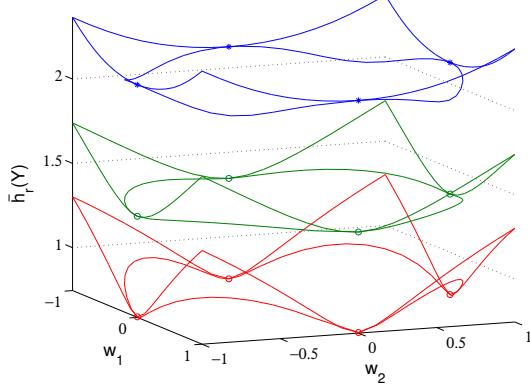
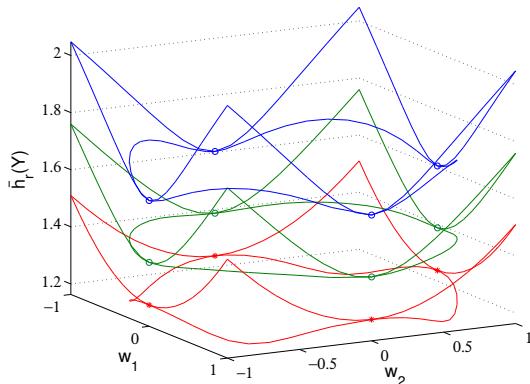
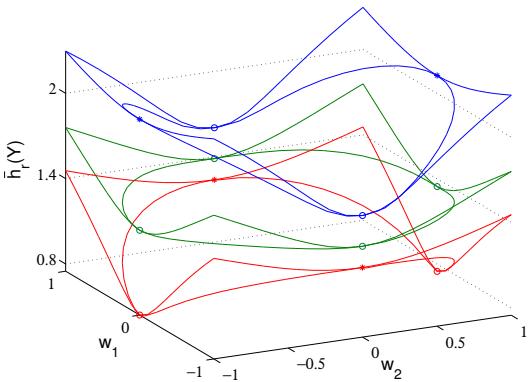
(a) Example of Fig. 2.3.(a): $\bar{h}_r(Y^\square)$ and $\bar{h}_r(Y^\circ)$ (b) Example of Fig. 2.3.(b): $\bar{h}_r(Y^\square)$ and $\bar{h}_r(Y^\circ)$ (c) Example of Fig. 2.4.: $\bar{h}_r(Y^\square)$ and $\bar{h}_r(Y^\circ)$

Figure 2.7. Rényi's entropy behaves differently depending on the constraint: the inequality (2.94) is satisfied under $\|\mathbf{w}\|_\infty = 1$ constrain (here, even if the sources have different densities), but not under the $\|\mathbf{w}\|_2 = 1$. The values of Rényi's exponent are $r = 0.2, r = 1, r = 5$ (remind that h_r is decreasing in r). The non-mixing local minima and maxima of $\bar{h}_r(Y)$ s.t. $\|\mathbf{w}\|_2 = 1$ are located by 'o' and '*' markers, respectively.

2.6 APPENDIX: PROOFS OF RESULTS OF THE CHAPTER

2.6.1 Proof of Corollary 6 (wording p. 50)

From Rem. 9, we are led to evaluate

$$f(\mathbf{B}) = \frac{1}{2} \log |\det(\mathbf{W}\mathbf{W}^T)| - \sum_{i=1}^P \log Q \left(\sum_{j=1}^K W_{ij} S_j \right) . \quad (2.97)$$

We know from Theorem 8 that the global maximum point of the above criterion is such that $\mathbf{W} = \mathbf{BA}$ is in the subset $\mathcal{W}^{P \times K}$ of $P \times K$ non-mixing matrices. For such matrices \mathbf{W} , there exists $j(i) \in \{1, \dots, K\}$ such that for all $i \in \{1, \dots, P\}$, $W_{ij} \neq 0$ if and only if $j = j(i)$. Then, $Q(\sum_{j=1}^K |W_{ij}| S_j)$ reduces to $|W_{ij(i)}| Q(S_{j(i)})$ from (Huber 1) and thus:

$$\begin{aligned} f(\mathbf{B}) &= \frac{1}{2} \log \prod_{i=1}^P W_{ij(i)}^2 - \sum_{i=1}^P \log |W_{ij(i)}| - \sum_{i=1}^P \log Q(S_{j(i)}) \\ &= - \sum_{i=1}^P \log Q(S_{j(i)}) . \end{aligned} \quad (2.98)$$

Clearly, this quantity is maximized when $\sum_{i=1}^P \log Q(S_{j(i)})$ is minimized that is, when the P sources with lowest value of Q have been recovered. Formally, if we define the set $\mathcal{J}_{\mathbf{W}}$ of the column indexes of \mathbf{W} containing a non-zero entry, $\mathcal{J}_{\mathbf{W}} \doteq \{j : \exists i \text{ s.t. } W_{ij} \neq 0\}$, then:

$$\begin{aligned} \underset{\mathbf{B}}{\operatorname{argmax}} f(\mathbf{B}) &= \{\mathbf{B} : \mathbf{BA} \in \mathcal{W}^{P \times K}, \#[\{j \in \mathcal{J}_{\mathbf{BA}} : j \leq P^m\}] = P^m, \\ &\quad \max \mathcal{J}_{\mathbf{BA}} \leq P^M\} , \end{aligned} \quad (2.99)$$

where $\#$ is the cardinal operator. Note that $\#[\mathcal{J}_{\mathbf{W}}] = P$ for any $\mathbf{W} \in \mathcal{W}^{P \times K}$.

□

2.6.2 Proof of Property 4 (wording p. 57)

Remark first that the usual convention for one-dimensional finite-support densities is to define them on the whole real line, and to set their values outside their support to zero: $p_X(x) = 0$ for all $x \in \mathbb{R} \setminus \Omega(X)$. Let us now turn to the proof of the score function properties.

The score function is zero-mean:

$$\begin{aligned}
 E[\psi_X(X)] &= \int -p_X(x)(\log p_X(x))' dx \\
 &= - \int p'_X(x) dx \\
 &= -[p_X(x)]_{-\infty}^{\infty} ,
 \end{aligned} \tag{2.100}$$

where we have used $\lim_{x \rightarrow -\infty} p_X(x) = \lim_{x \rightarrow \infty} p_X(x) = 0$ because p_X is a density, and thus integrable. This chain of equalities is equivalent to impose a weak regularity condition (used e.g. for proving Cramér -Rao bound [Cover and Thomas, 1991]) on p_X ensuring that one can interchange integration and differentiation, since:

$$E[\psi_X(X)] = - \int p'_X(x) dx = -d/dx \int p_X(x) dx = 0 . \tag{2.101}$$

On the other hand, by integration by parts,

$$\begin{aligned}
 E[X\psi_X(X)] &= - \int x p'_X(x) dx \\
 &= -[p_X(x)x]_{-\infty}^{\infty} + \int p_X(x) dx \\
 &= 1
 \end{aligned} \tag{2.102}$$

because for well-behaved functions (satisfying the regularity conditions), $p_X(x)$ goes faster to zero than $1/x$ as $x \rightarrow \infty$.

Finally, again by integration by parts, $E[\psi'_{S_j}(S_j)]$ can be rewritten as $E[\psi_{S_j}^2(S_j)]$, which is precisely Fisher's information [Cover and Thomas, 1991]: for any random variable X , we have

$$\begin{aligned}
 \int p_X(x) \left(\frac{-p'_X(x)}{p_X(x)} \right)' &= - \int p''_X(x) dx + \int p_X(x) \left(\frac{p'_X(x)}{p_X(x)} \right)^2 dx \\
 &= [p'_X(x)]_{-\infty}^{\infty} + E[\psi_X^2] ,
 \end{aligned} \tag{2.103}$$

where the first term is zero for well behaved densities.

To check that the inequality

$$E[\psi_X^2(X)] \text{Var}[X] \geq 1 \tag{2.104}$$

always holds and that the strict equality case holds true if and only if X is Gaussian, observe that

$$\begin{aligned} E[\psi_X^2(X)]Var[X] &= \frac{E[\psi_X^2(X)]Var[X]}{\left(E[X\psi_X(X)] - E[\psi_X(X)]E[X]\right)^2} \\ &= \frac{Var[\psi_X(X)]Var[X]}{Cov^2[X, \psi_X(X)]} \\ &= \frac{1}{Corr^2[X, \psi_X(X)]} \\ &\geq 1 \end{aligned}$$

note that the equality can be reached if and only if $Corr[X, \psi_X(X)] = 1$, i.e. when ψ_X is proportional to $X - E[X]$, which can only occur when p_X is Gaussian.

□

2.6.3 Proof of Theorem 11 (wording p. 58)

Assume that \mathbf{w} is close from \mathbf{e}_j so that its i -th component $\mathbf{w}(i)$ is close to 0 for $i \neq j$. Under the $\mathbf{w} \in \mathcal{S}(K)$ constraint, $\mathbf{w}(j) = \sqrt{1 - \sum_{i \neq j} \mathbf{w}(i)^2}$ and since $\sqrt{1 - x} = 1 - \frac{1}{2}x + o(x)$, one can write

$$\mathbf{w}(j) = 1 - \frac{1}{2} \sum_{i \neq j} \mathbf{w}(i)^2 + o\left(\sum_{i \neq j} \mathbf{w}(i)^2\right).$$

Thus, $\mathbf{w}\mathbf{S} = \mathbf{S}_j + \delta\mathbf{S}_j$ with

$$\delta\mathbf{S}_j = \sum_{i \neq j} \mathbf{w}(i)\mathbf{S}_i - \frac{1}{2} \left(\sum_{i \neq j} w_i^2 \right) \mathbf{S}_j + o\left(\sum_{i \neq j} \mathbf{w}(i)^2\right).$$

Therefore, applying (2.27) and dropping higher order terms, one gets that $h(\mathbf{w}\mathbf{S})$ equals

$$\begin{aligned} h(\mathbf{S}_j) + \left(\sum_{i \neq j} \mathbf{w}(i) \right) E[\psi_{\mathbf{S}_j}(\mathbf{S}_j)\mathbf{S}_i] - \frac{1}{2} \left(\sum_{i \neq j} \mathbf{w}(i)^2 \right) E[\psi'_{\mathbf{S}_j}(\mathbf{S}_j)\mathbf{S}_j] \\ + \frac{1}{2} \left\{ E \left[\text{Var} \left[\sum_{i \neq j} \mathbf{w}(i)^2 \mathbf{S}_i \mid \mathbf{S}_j \right] \psi'_{\mathbf{S}_j}(\mathbf{S}_j) \right] - \left(\sum_{i \neq j} \mathbf{w}(i) E[\mathbf{S}_i \mid \mathbf{S}_j] \right)^2 \right\} \\ + o\left(\sum_{i \neq j} \mathbf{w}(i)^2\right). \end{aligned}$$

Since the sources are mutually independent, any non-linear mapping of them is uncorrelated so that $E[\psi_{\mathbf{S}_j}(\mathbf{S}_j)\mathbf{S}_i] = 0$, for $i \neq j$. Furthermore $E[\mathbf{S}_i \mid \mathbf{S}_j] = E[\mathbf{S}_i] =$

0 for $i \neq j$, $E[\psi_{S_j}(S_j)S_j] = 1$, and $\text{Var}[\sum_{i \neq j} \mathbf{w}(i)S_i | S_j] = \text{Var}[\sum_{i \neq j} \mathbf{w}(i)S_i] = (\sum_{i \neq j} w_i^2)\sigma_S^2$ where σ_S^2 denotes the common variance of the sources. Therefore

$$h(\mathbf{w}S) = h(S_j) + \frac{1}{2} \left(\sum_{i \neq j} \mathbf{w}(i)^2 \right) \{ \sigma_S^2 E[\psi'_{S_j}(S_j)] - 1 \} + o\left(\sum_{i \neq j} \mathbf{w}(i)^2\right).$$

One concludes from Property 4 that for any non-Gaussian source $\sigma_S^2 E[\psi'_{S_j}(S_j)] > 1$, that is, $h(\mathbf{w}S) > h(S_j)$ for all \mathbf{w} sufficiently close to \mathbf{e}_j if S_j is non-Gaussian. Thus $h(\mathbf{w}S)$ reaches local non-mixing minima at $\mathbf{w} = \pm \mathbf{e}_j$ (since $h(-\mathbf{w}S) = h(\mathbf{w}S)$), as long as S_j is non-Gaussian. If S_j is Gaussian then $h(S_j)$ is a global maximum since Gaussian random variables have the highest entropy for a given variance. Equality (2.105) is of no use in this case, since the second term in this equality vanishes.

□

2.6.4 Proof of Lemma 6 (wording p. 60)

Suppose $\mu[\Omega(X)] = \mu[\bar{\Omega}(X)] > 0$ and $\mu[\Omega(Y)] = \mu[\bar{\Omega}(Y)] > 0$. This means that $\mu[\bar{\Omega}(X) \setminus \Omega(X)] = \mu[\bar{\Omega}(Y) \setminus \Omega(Y)] = 0$. Therefore, the sets $\Omega(X)$ and $\Omega(Y)$ can be expressed as

$$\begin{cases} \Omega(X) &= [\inf X, \sup X] \setminus \cup_{i=1}^{I'} \{x_i\} \\ \Omega(Y) &= [\inf Y, \sup Y] \setminus \cup_{j=1}^{J'} \{y_j\} \end{cases} \quad (2.105)$$

where x_i, y_i are isolated points. Then,

$$\begin{aligned} \mu[\Omega(X+Y)] &= \mu[\bar{\Omega}(X+Y)] \\ &= (\sup X + \sup Y) - (\inf X + \inf Y) \\ &= \mu[\Omega(X)] + \mu[\Omega(Y)], \end{aligned}$$

which yields the first result of the lemma.

To prove the second claim, suppose that $\Omega(X) = \cup_{i=1}^I \Omega_i(X)$ and $\Omega(Y) = \cup_{j=1}^J \Omega_j(Y)$. Further, $X^* = \cup_{i=1}^{I-1} [X_i^m, X_I^M] \setminus \cup_{i'=1}^{I'} \{x_{i'}\}$, $Y^* = \cup_{j=1}^{J-1} [Y_j^m, Y_J^M] \setminus \cup_{j'=1}^{J'} \{y_{j'}\}$ and $X = X^* \cup [X_I^m, X_I^M] \setminus \cup_{i^*=1}^{I^*} \{x_{i^*}\}$, $Y = Y^* \cup [Y_J^m, Y_J^M] \setminus \cup_{j^*=1}^{J^*} \{y_{j^*}\}$ where $X_i^m \leq X_i^M < X_{i+1}^m$, $Y_i^m \leq Y_i^M < Y_{i+1}^m$ and $X_I^m = X_{I-1}^M + \epsilon$, $\epsilon > 0$. We first assume that the right-most intervals constituting $\Omega(X)$ and $\Omega(Y)$ are not isolated points, that is have strictly positive measure: $\Delta_X \doteq X_I^M - X_I^m = \mu[\Omega_I(X)] > 0$ and $\Delta_Y \doteq Y_J^M - Y_J^m = \mu[\Omega_J(Y)] > 0$. Hence, we have:

$$\mu[\Omega(X+Y)] \geq \mu[\Omega(X^*+Y)] + \left\{ (Y_J^M + X_I^M) - \max(X_{I-1}^M + Y_J^m, Y_J^M + X_I^m) \right\},$$

where the term into brackets is a lower bound of the sub-volume of $\Omega(X+Y)$ due to the interval $[X_I^m, X_I^M]$; it can be rewritten as $\min\{\Delta_X + \epsilon, \Delta_X + \Delta_Y\}$.

Finally, having the Brunn-Minkowski inequality in mind, one gets:

$$\begin{aligned}\mu[\Omega(X+Y)] &\geq \mu[\Omega(X^*+Y)] + \min\{\Delta_X + \epsilon, \Delta_X + \Delta_Y\} \\ &\geq \mu[\Omega(X)] - \Delta_X + \mu[\Omega(Y)] + \min\{\Delta_X + \epsilon, \Delta_X + \Delta_Y\} \\ &> \mu[\Omega(X)] + \mu[\Omega(Y)].\end{aligned}$$

Suppose now that the right-most intervals might reduce to a single point, i.e. $X_i^m = X_i^M$ for $I - I_\star \leq i \leq I$, $Y_j^m = Y_j^M$ for $J - J_\star \leq j \leq J$ with $\min(I_\star, J_\star) \geq 1$. Because of the non-zero measure condition on the support sets, $I_\star < I$, $J_\star < J$. We rewrite the support set of the random variable as a union of a set having a strictly positive measure and of isolated points

$$\Omega(X) \doteq \Omega(X^*) \cup_{i=1}^{I^*} \{\xi_i\}, \quad \Omega(Y) \doteq \Omega(Y^*) \cup_{j=1}^{J^*} \{\zeta_j\},$$

such that there exists $\epsilon_x > 0$, $\epsilon_y > 0$ $[\sup \Omega(X^*) - \epsilon_x, \sup \Omega(X^*)] \subseteq \Omega(X^*)$, $[\sup \Omega(Y^*) - \epsilon_y, \sup \Omega(Y^*)] \subseteq \Omega(Y^*)$. In other words, each of the isolated points located on the right of the most right interval of $\Omega(X)$ (resp. $\Omega(Y)$) with strictly positive measure are relegated in the union $\cup_{i=1}^{I^*} \{\xi_i\}$ (resp. $\cup_{j=1}^{J^*} \{\zeta_j\}$). If such isolated points exist, one can always proceed to this trick because of the existence of the above “positive-measure” intervals: by hypothesis $\mu[\Omega(X)] > 0$, $\mu[\Omega(Y)] > 0$. As isolated points have zero-measure and do not affect the support measure of X and Y :

$$\mu[\Omega(X)] = \mu[\Omega^*(X)], \quad \mu[\Omega(Y)] = \mu[\Omega^*(Y)]$$

By contrast, they influence the support measure of $X+Y$. Indeed, $\Omega(X+Y)$ can be expressed as a union of subsets:

$$\underbrace{\{x+y : x \in \Omega^*(X), y \in \Omega^*(Y)\}}_{(a)} \cup_{i=1}^{I^*} \{y + \xi_i, y \in \Omega(Y)\} \cup_{j=1}^{J^*} \{x + \zeta_j, x \in \Omega(X)\}.$$

But the measure of the left-most set (a) is larger than or equal to the sum $\mu[\Omega^*(X)] + \mu[\Omega^*(Y)]$ (by the BMI) which precisely equals $\mu[\Omega(X)] + \mu[\Omega(Y)]$. On the other hand, at least one of the other sets cannot be totally included in (a). For instance, assuming that $\sup \Omega(X) \doteq \xi_S$, the term (b) is not totally contained in (a), and it can be shown that the remaining part “(b)\(a)” has a strictly positive measure. Indeed, because ξ_S is an isolated point, there exists $\epsilon > 0$ such that $\xi_S = \sup \Omega^*(X) + \epsilon$ and by definition of $\Omega^*(Y)$, $[\xi_S + \sup \Omega^*(Y) - \Delta, \xi_S + \sup \Omega^*(Y)]$ is included in $\Omega(X+Y)$ for all Δ satisfying $0 < \Delta \leq \epsilon_y$. But this interval has a strictly positive measure equal to Δ and is disjoint from (a) if $\xi_S + \sup \Omega^*(Y) - \Delta > \sup \Omega^*(X) + \sup \Omega^*(Y)$ that is if $\Delta < \epsilon$.

Hence, since for sufficiently small $\Delta > 0$ we have $\xi_S > \sup \Omega^*(Y) + \Delta$, it comes that

$$\mu[\Omega(X+Y)] \geq \mu[\Omega(X)] + \mu[\Omega(Y)] + \Delta,$$

for some $\Delta > 0$.

□

2.6.5 Proof of Theorem 14 (wording p. 62)

The proof of this theorem will be based on the two next propositions, and assumes $R(S_1) = R(S_2) = \dots = R(S_k) < R(S_{k+1})$.

Proposition 4 Let us define a $\mathbf{p} \in \mathcal{V}_K^\lambda$ vector respecting $\mathbf{p}(r) > 0$ for any $k < r \leq K$. Consider vector \mathbf{q} defined by:

$$\begin{cases} \mathbf{q}(r) = 0 \\ \mathbf{q}(k') = \sqrt{\mathbf{p}(k')^2 + \mathbf{p}(r)^2} \text{ with } 1 \leq k' \leq k \\ \mathbf{q}(j) = \mathbf{p}(j) \text{ for all } 1 \leq j \leq K, j \notin \{k', r\} \end{cases}, \quad (2.106)$$

Then, $\mathbf{q} \in \mathcal{V}_K^\lambda$ and $\tilde{\mathcal{C}}_R(\mathbf{q}) > \tilde{\mathcal{C}}_R(\mathbf{p})$, i.e. $\mathbf{p} \notin \{\mathbf{w} : \mathbf{w} = \arg \max_{\{\mathbf{w} \in \mathcal{V}_K^\lambda\}} \tilde{\mathcal{C}}_R(\mathbf{w})\}$.

Proof: It is trivial to show that $\mathbf{q} \in \mathcal{V}_K^\lambda$. On the other hand, we have

$$\mathbf{p}(r)^2 R^2(S_{k'}) < \mathbf{p}(r)^2 R^2(S_r),$$

and:

$$\begin{aligned} \mathbf{p}(k')^2 R^2(S_{k'}) + \mathbf{p}(r)^2 R^2(S_{k'}) &< \mathbf{p}(k')^2 R^2(S_{k'}) + \mathbf{p}(r)^2 R(S_r) \\ &\quad + \underbrace{2\mathbf{p}(k')\mathbf{p}(r)R(S_{k'})R(S_r)}_{\geq 0} \\ R(S_{k'})\sqrt{\mathbf{p}(k')^2 + \mathbf{p}(r)^2} &< \mathbf{p}(k')R(S_{k'}) + \mathbf{p}(r)R(S_r) \end{aligned} \quad (2.107)$$

Hence, it results from the definition of \mathbf{q} that $-\tilde{\mathcal{C}}_R(\mathbf{q}) < -\tilde{\mathcal{C}}_R(\mathbf{p})$ and thus $\tilde{\mathcal{C}}_R(\mathbf{q}) > \tilde{\mathcal{C}}_R(\mathbf{p})$. \square

Proposition 5 For any $\mathbf{p} \in \mathcal{V}_K^\lambda$ vector satisfying $\mathbf{p}(j) = 0$ for all $k < j \leq K$, then $\tilde{\mathcal{C}}_R(\mathbf{p}) \leq \tilde{\mathcal{C}}_R(\lambda e_j)$, $1 \leq j \leq k$ with equality if and only if $\mathbf{p} \in \{\lambda e_1, \dots, \lambda e_k\}$

Proof: If $\mathbf{p}(j) = 0$ for all $j > k$, then, because $\mathbf{p} \in \mathcal{V}_K^\lambda$, there must exist $r \leq k$ such that $\mathbf{p}(r) > 0$. On the other hand, for any $1 \leq r \neq r' \leq k$, we know that $\mathbf{p}(r') \geq 0$. Hence, by definition of k :

$$\mathbf{p}(r)R(S_r) + \mathbf{p}(r')R(S_{r'}) = (\mathbf{p}(r) + \mathbf{p}(r'))R(S_r). \quad (2.108)$$

Let us define \mathbf{q} by $\mathbf{q}(j) = \mathbf{p}(j)$ for $j \notin \{r, r'\}$, $\mathbf{q}(r) = \sqrt{\mathbf{p}(r)^2 + \mathbf{p}(r')^2}$ and $\mathbf{q}(r') = 0$. Then, it is straightforward to show that $\mathbf{q} \in \mathcal{V}_K^\lambda$, and that $\tilde{\mathcal{C}}_R(\mathbf{q}) \geq \tilde{\mathcal{C}}_R(\mathbf{p})$ with equality if and only if $\mathbf{p}(r') = 0$. To prove the last claim, remark that:

$$\sqrt{\mathbf{p}(r)^2 + \mathbf{p}(r')^2} \leq \mathbf{p}(r) + \mathbf{p}(r'). \quad (2.109)$$

with equality only when $\mathbf{p}(r') = 0$. Hence, by iterating this result setting $\mathbf{p} \leftarrow \mathbf{q}$, if such a \mathbf{p} vector has at least two strictly positive elements, then $\tilde{\mathcal{C}}_R(\mathbf{p}) <$

$\tilde{\mathcal{C}}_R(\lambda \mathbf{e}_j)$, with $1 \leq j \leq k$. On the other hand, it is easy to see that if a \mathbf{p} vector satisfying $\mathbf{p}(k+1) = \dots = \mathbf{p}(K) = 0$ and $\mathbf{p} \in \mathcal{V}_K^\lambda$ has a single non-zero entry, then $\mathbf{p} \in \{\lambda \cdot \mathbf{e}_1, \dots, \lambda \cdot \mathbf{e}_k\}$. This concludes the proof of the proposition. By iterating Proposition 4, for any vector $\mathbf{p} \in \mathcal{V}_K^\lambda$ such that there exists $k < r \leq K$ with $\mathbf{p}(r) > 0$ there exists another vector $\mathbf{q} \in \mathcal{V}_K^\lambda$, respecting $\mathbf{q}(j) = 0$ for all $k < j \leq K$ satisfying $\tilde{\mathcal{C}}_R(\mathbf{q}) > \tilde{\mathcal{C}}_R(\mathbf{p})$. On the other hand, Proposition 5 shows that among all those \mathbf{q} vectors, only $\mathbf{q} \in \{\lambda \cdot \mathbf{e}_1, \dots, \lambda \cdot \mathbf{e}_k\}$ can maximize globally function $\tilde{\mathcal{C}}_R$ subjected to $\mathbf{q} \in \mathcal{V}_K^\lambda$.

□

2.6.6 Proof of Theorem 15 (wording p. 63)

Suppose that $\hat{\mathbf{e}}_i \in \mathcal{V}_K^1$ is a vector close to \mathbf{e}_i , in the sense that $\hat{\mathbf{e}}_i = \mathbf{e}_i + \delta \mathbf{e}_i$ where $\delta \mathbf{e}_i$ is a “small” vector. Obviously, $\delta \mathbf{e}_i(i) < 0$ and $\delta \mathbf{e}_i(j) \geq 0$, for $j \neq i$. We note $\delta \mathbf{e}_i(i) = -\epsilon$ where $\epsilon > 0$. By the $\|\hat{\mathbf{e}}_i\| = 1$ constraint, it comes that:

$$\begin{aligned} \sum_{j \neq i} \hat{\mathbf{e}}_i(j)^2 &= 1 - \hat{\mathbf{e}}_i(i)^2 \\ &= 1 - (1 - \epsilon)^2 . \end{aligned} \quad (2.110)$$

On the other hand, by Eq. (2.50):

$$\Delta \tilde{\mathcal{C}}_R(\mathbf{e}_i, \hat{\mathbf{e}}_i) = \underbrace{(1 - \epsilon)R(\mathbf{S}_i) + \sum_{j \neq i} \hat{\mathbf{e}}_i(j)R(\mathbf{S}_j)}_{-\tilde{\mathcal{C}}_R(\hat{\mathbf{e}}_i)} - \underbrace{R(\mathbf{S}_i)}_{-\tilde{\mathcal{C}}_R(\mathbf{e}_i)} . \quad (2.111)$$

Hence, Theorem 15 will be proven if

$$\sum_{j \neq i} \hat{\mathbf{e}}_i(j)R(\mathbf{S}_j) > \epsilon R(\mathbf{S}_i) . \quad (2.112)$$

Let us denote the norm of $\hat{\mathbf{e}}_i$ s.t. $j \neq i$ vector by:

$$\lambda' \doteq \sqrt{\sum_{j \neq i} \hat{\mathbf{e}}_i(j)^2} . \quad (2.113)$$

By Eq. (2.110), $\lambda' = \sqrt{1 - (1 - \epsilon)^2}$. Hence, by using Theorem 14 with $\mathbf{w} = [\hat{\mathbf{e}}_i(1), \dots, \hat{\mathbf{e}}_i(i-1), \hat{\mathbf{e}}_i(i+1), \dots, \hat{\mathbf{e}}_i(K)]$ and $\mathbf{w} \in \mathcal{V}_{K-1}^{\lambda'}$, we find $\tilde{\mathcal{C}}_R(\mathbf{w}) \leq \tilde{\mathcal{C}}_R(\lambda' \mathbf{e}_r)$, where $r = \arg \min_{j \neq i} \{R(\mathbf{S}_j)\}$. In other words, the following inequality holds:

$$\sum_{j \neq i} \hat{\mathbf{e}}_i(j)R(\mathbf{S}_j) \geq \underbrace{\sqrt{1 - (1 - \epsilon)^2}}_{\lambda'} R(\mathbf{S}_r) . \quad (2.114)$$

Then, having Eq. (2.112) in mind, a sufficient condition to prove Theorem 2 is to check that the following inequality holds for any sufficiently small $\epsilon > 0$:

$$\lambda' R(\mathbf{S}_r) > \epsilon R(\mathbf{S}_i) \text{ with } r \neq i . \quad (2.115)$$

By transitivity, the previous inequality holds when:

$$\begin{aligned} \sqrt{2\epsilon - \epsilon^2} R(\mathbf{S}_r) &> \epsilon R(\mathbf{S}_i) \\ (2\epsilon - \epsilon^2)R^2(\mathbf{S}_r) &> \epsilon^2 R^2(\mathbf{S}_i) . \end{aligned} \quad (2.116)$$

Hence, if $\epsilon[2R^2(\mathbf{S}_r) - \epsilon(R^2(\mathbf{S}_r) + R^2(\mathbf{S}_i))] > 0$ holds for any sufficiently small $\epsilon > 0$, then Eq. (2.112) is fulfilled.

The last inequality is satisfied for all $0 < \epsilon < \frac{2R^2(\mathbf{S}_r)}{R^2(\mathbf{S}_i) + R^2(\mathbf{S}_r)}$. This result concludes the proof: $\Delta\tilde{\mathcal{C}}_R(\mathbf{e}_i, \hat{\mathbf{e}}_i) > 0$ for all sufficiently small $\epsilon > 0$.

□

2.6.7 Proof of Theorem 16 (wording p. 64)

The proof of this theorem results from an adaptation to $P \leq K$ of the proof of *Proposition 3* presented in [Pham, 2000]. As in this proof, in order to show that any matrix in $\mathcal{W}^{P \times K}$ is a local maximum point of $\tilde{\mathcal{C}}_R$ it is sufficient to prove that for a small increment $\delta\mathbf{W}$ of $\mathbf{W} \in \mathcal{W}^{P \times K}$, the quantity

$$\sum_{i=1}^P \left\{ \log \left[\sum_{j=1}^K |W_{ij} + \delta W_{ij}| R(\mathbf{S}_j) \right] - \log \left[\sum_{j=1}^K |W_{ij}| R(\mathbf{S}_j) \right] \right\} , \quad (2.117)$$

W_{ij} and δW_{ij} denoting the general element of \mathbf{W} and of $\delta\mathbf{W}$, is larger or equal than $\frac{1}{2}\{\log \det[(\mathbf{W} + \delta\mathbf{W})(\mathbf{W} + \delta\mathbf{W})^T] - \log \det(\mathbf{W}\mathbf{W}^T)\}$, up to first order in $\delta\mathbf{W}$. But since $\mathbf{W} \in \mathcal{W}^{P \times K}$, there exists distinct indexes $j(1), \dots, j(P)$ such that for $i = 1, \dots, P$, $W_{ij} \neq 0$ if and only if $j = j(i)$. Thus (2.117) reduces to

$$\sum_{i=1}^P \log \left[\left| 1 + \frac{\delta W_{ij(i)}}{W_{ij(i)}} \right| + \frac{\sum_{j \neq j(i)}^K |\delta W_{ij}| R(\mathbf{S}_j)}{|W_{ij(i)}| R(\mathbf{S}_{j(i)})} \right] ,$$

which, for $|\delta W_{ij(i)}| < |W_{ij(i)}|$, equals

$$\sum_{i=1}^P \left[\frac{\delta W_{ij(i)}}{W_{ij(i)}} + \frac{\sum_{j \neq j(i)}^K |\delta W_{ij}| R(\mathbf{S}_j)}{|W_{ij(i)}| R(\mathbf{S}_{j(i)})} \right] + O(\|\delta\mathbf{W}\|^2) .$$

On the other hand, the first order Taylor expansion of a multivariate function $f : \mathbb{R}^{P \times K} \rightarrow \mathbb{R}$ is

$$f(\mathbf{W} + \delta\mathbf{W}) = f(\mathbf{W}) + \langle \nabla f, \delta\mathbf{W} \rangle + \dots \quad (2.118)$$

where ∇f is the gradient of f and $\langle \cdot, \cdot \rangle$ denotes the dot product. But, as it will be shown in Lemma 21 (see Chapter 3, Section 3.4.4), we have

$$\frac{\partial \log |\det(\mathbf{W}\mathbf{W}^T)|}{\partial W_{ij}} = 2[(\mathbf{W}^+)^T]_{ij} = 2[\mathbf{W}^+]_{ji} \quad (2.119)$$

implying

$$\langle \nabla f, \delta \mathbf{W} \rangle = 2 \sum_{i,j} [(\mathbf{W}^+)^T]_{ij} \delta W_{ij} = 2 \text{Tr} [\mathbf{W}^+ \delta \mathbf{W}] . \quad (2.120)$$

Note that due to the special form of \mathbf{W} , $\mathbf{W}\mathbf{W}^T$ is diagonal with i -th diagonal equal to $W_{ij(i)}^2$. Thus, $\mathbf{W}^+ = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}$ has its ji element equal to 0 if $j \neq j(i)$ and to $1/W_{ij(i)}$ otherwise. Therefore

$$\frac{1}{2} \{ \log \det[(\mathbf{W} + \delta \mathbf{W})(\mathbf{W} + \delta \mathbf{W})^T] - \log \det(\mathbf{W}\mathbf{W}^T) \} = \sum_{i=1}^P \frac{\delta W_{ij(i)}}{W_{ij(i)}} + O(\|\delta \mathbf{W}\|^2) .$$

It follows that (2.117) is greater than the above left hand side, up to a term of order $O(\|\delta \mathbf{W}\|^2)$.

□

2.6.8 Convolution of Gaussian kernels (wording p. 67)

In this section, the last equality of Eq. (2.58) is proved. This is a well-known result but we were not able to find the proof. Therefore, we propose here a personal development, without guarantee that the approach is original. Let us first prove that the convolution of two centered Gaussian kernels $\phi_{\sigma_i}(x)$ and $\phi_{\sigma_j}(x)$ is equal to a Gaussian kernel with standard deviation equal to $\sigma_{ij} = \sqrt{\sigma_i^2 + \sigma_j^2}$:

$$\phi_{\sigma_i}(x) * \phi_{\sigma_j}(x) \doteq \int_{-\infty}^{+\infty} \phi_{\sigma_i}(\tau) \phi_{\sigma_j}(x - \tau) d\tau = \phi_{\sqrt{\sigma_i^2 + \sigma_j^2}}(x) . \quad (2.121)$$

Observe that

$$\begin{aligned} \phi_{\sigma_i}(x) * \phi_{\sigma_j}(x) &= \frac{1}{2\pi\sigma_i\sigma_j} \int e^{-\frac{\tau^2}{2\sigma_i^2}} e^{-\frac{(x-\tau)^2}{2\sigma_j^2}} d\tau \\ &= \frac{1}{2\pi\sigma_i\sigma_j} \int_{-\infty}^{+\infty} e^{-\frac{\tau^2}{2\sigma_i^2}} e^{-\frac{(x-\tau)^2}{2\sigma_j^2}} \underbrace{e^{-\frac{x^2}{2(\sigma_i^2 + \sigma_j^2)}} e^{\frac{x^2}{2(\sigma_i^2 + \sigma_j^2)}}}_{=1} d\tau \\ &= \underbrace{\frac{1}{2\pi\sigma_i\sigma_j} e^{-\frac{x^2}{2(\sigma_i^2 + \sigma_j^2)}}}_{\xi} \int_{-\infty}^{+\infty} e^{-\frac{\tau^2}{2\sigma_i^2}} e^{-\frac{(x-\tau)^2}{2\sigma_j^2}} e^{\frac{x^2}{2(\sigma_i^2 + \sigma_j^2)}} d\tau \\ &= \xi \int_{-\infty}^{+\infty} e^{-\frac{(\tau(\sigma_i^2 + \sigma_j^2) - x\sigma_i^2)^2}{2\sigma_i^2\sigma_j^2(\sigma_i^2 + \sigma_j^2)}} d\tau \\ &= \xi \int_{-\infty}^{+\infty} e^{-\frac{(\frac{\tau(\sigma_i^2 + \sigma_j^2)}{\sigma_i\sigma_j} - x\frac{\sigma_i}{\sigma_j})^2}{2(\sigma_i^2 + \sigma_j^2)}} d\tau \end{aligned} \quad (2.122)$$

Let us set $\zeta \doteq \frac{\tau(\sigma_i^2 + \sigma_j^2)}{\sigma_i \sigma_j} - x \frac{\sigma_i}{\sigma_j}$, , then $d\tau = \frac{\sigma_i \sigma_j}{\sigma_i^2 + \sigma_j^2} d\zeta$ and we get from the above chain of equalities:

$$\phi_{\sigma_i}(x) * \phi_{\sigma_j}(x) = \xi \frac{\sigma_i \sigma_j}{\sigma_i^2 + \sigma_j^2} \int_{-\infty}^{+\infty} e^{-\frac{\zeta^2}{2(\sigma_i^2 + \sigma_j^2)}} d\zeta \quad (2.123)$$

$$= \xi \frac{\sigma_i \sigma_j}{\sigma_i^2 + \sigma_j^2} \sqrt{2\pi} \sqrt{\sigma_i^2 + \sigma_j^2} \quad (2.124)$$

$$= \phi_{\sqrt{\sigma_i^2 + \sigma_j^2}}(x) , \quad (2.125)$$

which shows that convoluting two centered Gaussian kernels with standard deviation of, respectively, σ_i and σ_j yields another centered Gaussian kernel of standard deviation $\sqrt{\sigma_i^2 + \sigma_j^2}$.

Let us now prove that the integral of the product of two Gaussian functions of means μ_i and μ_j and variances σ_i and σ_j equals $\phi_{\sqrt{\sigma_i^2 + \sigma_j^2}}(\mu_j - \mu_i)$:

$$\int_{-\infty}^{+\infty} \phi_{\sigma_i}(x - \mu_i) \phi_{\sigma_j}(x - \mu_j) dx = \phi_{\sqrt{\sigma_i^2 + \sigma_j^2}}(\mu_j - \mu_i) . \quad (2.126)$$

By setting $\tau \doteq x - \mu_i$, we have :

$$x - \mu_j = x - \mu_i + (\mu_i - \mu_j) \quad (2.127)$$

$$= \tau + \underbrace{\mu_i - \mu_j}_{\doteq -s} \quad (2.128)$$

Then, by substitution, we find that

$$\int_{-\infty}^{+\infty} \phi_{\sigma_i}(x - \mu_i) \phi_{\sigma_j}(x - \mu_j) dx = \int \phi_{\sigma_i}(\tau) \phi_{\sigma_j}(s - \tau) d\tau \quad (2.129)$$

$$= \phi_{\sigma_i}(s) * \phi_{\sigma_j}(s) \quad (2.130)$$

$$= \phi_{\sqrt{\sigma_i^2 + \sigma_j^2}}(\mu_j - \mu_i) \quad (2.131)$$

□

and this concludes the proof of Eq. (2.126).

2.6.9 Proof of Lemma 7 (wording p. 71)

Let us write $\mathbf{w}_i = \mathbf{b}_i \mathbf{A}$, as usual. Then, $h_r(\mathbf{b}_i \mathbf{X}) = h_r(\mathbf{w}_i \mathbf{S})$, and for a small increment $[\delta_1, \dots, \delta_K]$ of \mathbf{w}_i :

$$-h_r(\mathbf{w}_i \mathbf{S} + [\delta_1, \dots, \delta_K] \mathbf{S}) = -h_r(\mathbf{w}_i \mathbf{S}) + \sum_{k=1}^K \delta_k E[\psi_{\mathbf{w}_i \mathbf{S}, r}(\mathbf{w}_i \mathbf{S}) S_k]$$

$$+o\left[\left(\sum_{k=1}^K \delta_k^2\right)^{1/2}\right]. \quad (2.132)$$

Further, using Eq. (2.19), the output unit-variance constraint gives $\|\mathbf{w}_i\| = \mathbf{w}_i \mathbf{w}_i^T = 1$, which yields $[\delta_1, \dots, \delta_K] \mathbf{w}_i^T = o[\sum_{k=1}^K \delta_k^2]$, i.e. $[\delta_1, \dots, \delta_K] \mathbf{w}_i^T$ is $o[(\sum_{k=1}^K \delta_k^2)^{1/2}]$. Thus if $\mathbf{w}_i = \pm \mathbf{e}_j$, then $\delta_j = o[(\sum_{k=1}^K \delta_k^2)^{1/2}]$ and

$$-\mathbf{h}_r(\pm \mathbf{S}_j + [\delta_1, \dots, \delta_K] \mathbf{S}) = -\mathbf{h}_r(\pm \mathbf{S}_j) + o\left[\left(\sum_{k=1}^K \delta_k^2\right)^{1/2}\right] \quad (2.133)$$

meaning that the scale-invariant functional $\tilde{\mathcal{C}}_{h_r}(\mathbf{w})$ admits a stationary point at $\pm \mathbf{e}_j$ or, equivalently, that $\mathcal{C}_{h_r}(\mathbf{b})$ is stationary when $\mathbf{bA} \in \{\pm \mathbf{e}_1, \dots, \pm \mathbf{e}_K\}$ (i.e. when the unit norm vector \mathbf{bA} satisfies $\|\mathbf{bA}\|_\infty = 1$).

□

2.6.10 Proof of Lemma 8 (wording p. 71)

Let us develop Rényi's entropy up to second order around $\mathbf{w}_j = \mathbf{e}_j$ (observe that Rényi's entropy is not sensitive to the sign of the random variable). We find

$$\begin{aligned} -\mathbf{h}_r(\pm \mathbf{S}_j + [\delta_1, \dots, \delta_K] \mathbf{S}) &= -\mathbf{h}_r[(\delta_j \pm 1) \mathbf{S}_j] - \frac{1}{2} \sum_{k, 1 \leq k \neq j \leq K} \delta_k^2 J_r(\pm \mathbf{S}_j) \\ &\quad + o\left(\sum_{k, 1 \leq k \neq j \leq K} \delta_k^2\right) \end{aligned} \quad (2.134)$$

But $J_r(\mathbf{S}_j) = J_r(-\mathbf{S}_j)$ and $\mathbf{h}_r[(\delta_j \pm 1) \mathbf{S}_j] = \mathbf{h}_r(\pm \mathbf{S}_j) + \log |1 \pm \delta_j|$ and further the constraint $\|\mathbf{w}_i\| = 1$ yields $|1 \pm \delta_i|^2 = 1 - \sum_{k \neq j} \delta_k^2$. Therefore

$$\begin{aligned} -\mathbf{h}_r(\pm \mathbf{S}_j + [\delta_1, \dots, \delta_K] \mathbf{S}) &= -\mathbf{h}_r(\pm \mathbf{S}_j) - \frac{1}{2} \sum_{k, 1 \leq k \neq j \leq K} \delta_k^2 [J_r(\mathbf{S}_j) - 1] \\ &\quad + o\left(\sum_{k, 1 \leq k \neq j \leq K} \delta_k^2\right). \end{aligned} \quad (2.135)$$

The above result shows that a necessary condition for the function $-\mathbf{h}_r(\mathbf{wS})$ over the set $\mathcal{S}(K)$ to admit a local maximum at $\pm \mathbf{e}_j$ is that $J_r(\mathbf{S}_j) \geq 1$ and a sufficient condition is that this inequality is strict. Since the sources have been assumed to have unit variance, one can write these conditions as $J_r(\mathbf{S}_j) \text{Var}[\mathbf{S}_j] \geq 1$ and $J_r(\mathbf{S}_j) \text{Var}[\mathbf{S}_j] > 1$, which are then independent of the source variance.

□

2.6.11 Proof of Lemma 9 (wording p. 72)

Let us evaluate the Rényi entropy-based simultaneous BSS criterion around \mathbf{B} , i.e. at $\mathbf{B} + \mathcal{E}\mathbf{B}$ where \mathcal{E} is a “small” matrix:

$$\mathcal{C}_{h_r}(\mathbf{B} + \mathcal{E}\mathbf{B}) = \log |\det(\mathbf{B} + \mathcal{E}\mathbf{B})| - \sum_{i=1}^K h_r[(\mathbf{B} + \mathcal{E}\mathbf{B})_i \mathbf{X}] . \quad (2.136)$$

Further, it will be shown that $\log |\det(\mathbf{B} + \mathcal{E}\mathbf{B})| = \log |\det \mathbf{B}| + \text{Tr}(\mathcal{E}) - 1/2\text{Tr}(\mathcal{E}^2) + o(\|\mathcal{E}\|^2)$ (see Eq. (3.14) and the associated proof in Section 3.8.1, p. 167). If \mathcal{E}_{ij} denotes the general element of \mathcal{E} and if we express the small vector $[\delta_1, \dots, \delta_K]$ in Eq. (2.132) in the basis spanned by the columns of the regular matrix $\mathbf{W} = \mathbf{B}\mathbf{A}$ with coefficients given by a i -th row of \mathcal{E} (i.e. with $\delta_j = \sum_{k=1}^K \mathcal{E}_{ik} W_{kj}$), one gets

$$\mathcal{C}_{h_r}(\mathbf{B} + \mathcal{E}\mathbf{B}) = \mathcal{C}_{h_r}(\mathbf{B}) - \sum_{i \neq j} \sum \mathcal{E}_{ij} \mathbb{E}[\psi_{Y_i, r}(Y_i) Y_j] + o(\|\mathcal{E}\|) . \quad (2.137)$$

Since, for any pair of functions f, g $\mathbb{E}[f(S_i)g(S_j)] = 0$ if $i \neq j$, it is seen that the criterion $\mathcal{C}_{h_r}(\mathbf{B})$ is stationary when the Y_k coincide with the sources, up to a scale factor.

□

2.6.12 Proof of Lemma 10 (wording p. 72)

From the above result, if the components Y_k of $\mathbf{Y} = \mathbf{B}\mathbf{X}$ are proportional to distinct sources, hence independent, the expansion of $\mathcal{C}_{h_r}(\mathbf{B} + \mathcal{E}\mathbf{B})$ takes the form

$$\mathcal{C}_{h_r}(\mathbf{B} + \mathcal{E}\mathbf{B}) = \mathcal{C}_{h_r}(\mathbf{B}) - \frac{1}{2} \sum_{1 \leq i \neq j \leq K} [\mathcal{E}_{ij}^2 J_r(Y_i) \text{Var}[Y_j] + \mathcal{E}_{ij} \mathcal{E}_{ji}] . \quad (2.138)$$

The last sum is a quadratic form associated with the symmetric block diagonal matrix, with 2×2 blocks:

$$\mathbf{J}_r^{i,j} \doteq \begin{bmatrix} J_r(Y_i) \text{Var}[Y_j] & 1 \\ 1 & J_r(Y_j) \text{Var}[Y_i] \end{bmatrix} , \quad (2.139)$$

that is

$$\sum_{1 \leq i \neq j \leq K} [\mathcal{E}_{ij}^2 J_r(Y_i) \text{Var}[Y_j] + \mathcal{E}_{ij} \mathcal{E}_{ji}] = \sum_{1 \leq i < j \leq K} [\mathcal{E}_{ij} \mathcal{E}_{ji}] \mathbf{J}_r^{i,j} [\mathcal{E}_{ij} \mathcal{E}_{ji}]^\top . \quad (2.140)$$

Thus, in order that the criterion $\mathcal{C}_{h_r}(\mathbf{B})$ attain a local maximum at the point $\mathbf{B} \sim \mathbf{A}^{-1}$, which is the same as the Y_i be proportional to distinct sources, it is necessary that the $\mathbf{J}_r^{i,j}$ matrices in Eq. (2.139) be positive semi-definite and it is

sufficient that they are positive definite. But a necessary condition for the $\mathbf{J}_r^{i,j}$ to be positive definite is to have a positive determinant, and a necessary condition becomes $J_r(Y_i)\text{Var}[Y_i]J_r(Y_j)\text{Var}[Y_j] \geq 1, \forall i \neq j$. The sufficient condition is that the above inequality is strict.

Note that the product $J_r(Y_i)\text{Var}[Y_i]$ is scale invariant: it is unchanged when Y_i is multiplied by a constant factor. Thus, in the case where the sources have the same density, the above necessary condition reduces to $J_r(S)\text{Var}[S] \geq 1$ where S is a random variable with density equal to the common density of the sources. The sufficient condition is $J_r(S)\text{Var}[S] > 1$.

□

CHAPTER 3

DISCRIMINACY OF ENTROPIC CONTRASTS

ANALYSIS OF THE MIXING MAXIMA

Abstract. Chapter 2 addressed the possible existence of local non-mixing maximum points of criteria based on the (extended form of) Rényi’s entropy. Furthermore, as discussed in Section 1.8, adaptive optimization techniques similar to those given in Section 1.6 may lead to any local maximum point, mixing or not. Therefore, if mixing local maximum points exist, the algorithm may be stuck in such a solution, which is actually a spurious solution. Some methods exist to look for global maximum points like e.g. simulated annealing (which basically consists in maximizing *powers* of contrast functions to attenuate the local maxima compared to the global maximum); by the contrast function definition, these global maximum points are necessarily non-mixing. However, *local* maximum points can also be non-mixing (in both deflation and partial separation schemes), and these points will not be recovered by using such optimization techniques. Actually, what we would like to do is “simply” to converge to any non-mixing point (i.e. to any local maximum point corresponding to the extraction of – a subset of – the sources).

To that aim, one needs to know in advance if mixing maximum points exist. If such points do not exist, we know that the solution provided by the iterative optimization technique will give a (possibly local) maximum point, but this point shall correspond to an acceptable solution of the BSS problem. On the contrary, this is no more true if mixing maxima exist: we have no guarantee that the solution found is acceptable. The goal of this chapter is to deal with this question for the contrast functions provided in Chapter 2.

Contribution. As in Chapter 2, the original results about the mixing maxima of the entropic contrasts (namely: the Shannon entropy-based, the support-based and the range-based contrasts) are summarized. Next, the mathematical tools that have been developed in order to perform the above analysis are listed. Finally, intuitive justifications of some phenomena are given.

- Results about the local and global optima of entropic criteria
 - **Shannon’s entropy**-based BSS contrast functions was proved to suffer from mixing optima based on experimental results when the source densities are multimodal; but those simulations always involved entropy approximation, so that it was unclear whether such mixing maxima exist in the exact Shannon’s entropy-based contrast. This is rigorously proved here when the source densities are multimodal enough. Mutual information is also proved to have such spurious optima. Hence, Shannon’s entropy-based contrasts are not discriminant contrasts in deflation and simultaneous separation schemes (and consequently for partial separation scheme, too)
 - **Hartley’s entropy**-based BSS contrast functions are proved to suffer from the same drawback when (shortly) the source supports are non-convex; this can be seen to be related to the “multimodality” concept when only the support is considered.
 - **Extended Hartley’s entropy**-based (i.e. range-based) BSS contrast functions are proved to be, on the contrary, discriminant contrast functions in the three extraction schemes, and thus even if a prewhitening step is not performed. To our knowledge, this is the single BSS criterion that is proved to benefit from the discriminacy property in the three extraction schemes (and in addition when no prewhitening is used), so far.
- Tools and other results
 - The Taylor expansion of Shannon’s entropy was useful to give counter-examples showing that the related deflation and simultaneous contrast functions are not discriminant. However, this could be shown for a limited class of source densities (two i.i.d. and symmetric sources). Therefore, another technique is presented, based on entropy approximation, exploiting the multimodal nature of the source pdfs. It allows us to extend the results to pairs of sources that have different and asymmetric multimodal densities. Error bounds are also provided.
 - The output range is proved to be a g-convex functional under a fixed-variance constraint. This useful (but unrecognized) result implies the discriminacy of the deflation range-based contrast.

- Intuitive developments

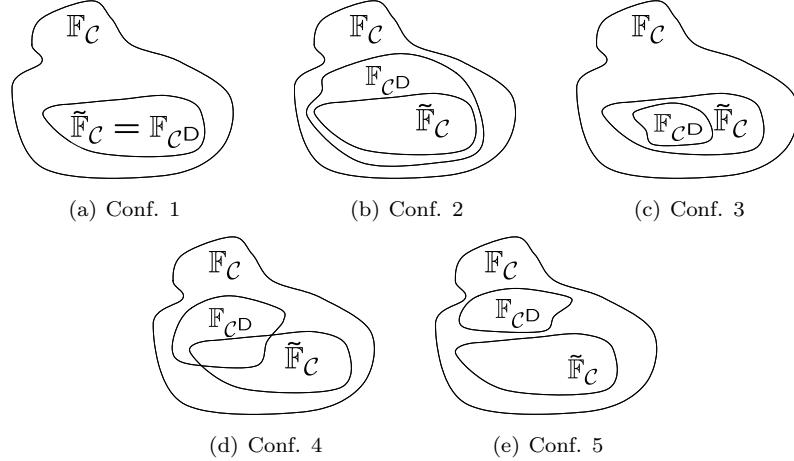
- We propose an intuitive (informal) justification explaining the existence of mixing maxima that emphasizes the specific nature of the multimodal source densities and, point out the relationship between the modality of a random variable and its entropy in our BSS framework.
- It is intuitively explained why cumulant-based criteria (like absolute kurtosis) do not suffer from this drawback.

In this chapter, all the proofs are original. Some of them result from joint work with D.-T. Pham. Part of the work presented in this chapter was or will be published in the following papers (see Appendix B): JA1, JP1, JP2, JP3, JA1, ICB1, ICP7, ICP8 (results about Shannon’s entropy) JA2, JA3, JTBS1, ICB6, ICP6, ICP10(results about the support and the range) JS1, ICTBS1 (results about Rényi entropies).

Organization of the chapter. In Section 3.2, the possible existence of mixing minima of Shannon’s entropy is analyzed. An informal justification and two rigorous approaches are used to show that this may not be the case when the sources have multimodal densities. As it is known that there is no mixing optimum in some cumulant-based criteria, we try to sketch an intuitive reasoning explaining the origin of this difference. The mixing extrema of the general form of Rényi’s entropies are not investigated as they do not lead to contrast functions. Then the range criterion is analyzed in Section 3.4, from various viewpoints and for the three extraction schemes. Section 3.5 shows that the discriminacy property is shared by the range but not by the support, which behaves similarly to the Shannon entropy.

3.1 CONCEPT DEFINITION AND TERMINOLOGY JUSTIFICATION

Contrast functions that are free from mixing maximum points will be called *discriminant*, in the sense that the characterization of their local maxima may render us able to distinguish between spurious and optimal solutions as the set of the first kind of solutions is empty. On the contrary, for the other contrast functions, we cannot guarantee that the demixing matrix found via adaptive maximization is (sub)PD-equivalent to the inverse of the mixing matrix. As an example, in the partial separation scheme, an output mutual information close to zero does not tell us anything about the quality of the separation, as explained in the introductory paragraph of Section 1.4. Basically, Comon’s terminology expressed that a criterion is a contrast function if (among others), when $\mathbf{BA} \sim \mathbf{I}_K$, then the criterion reaches its maximum value. It was a discriminating contrast if the global maximum occurs only when \mathbf{B} is of the form $\mathbf{BA} \sim \mathbf{I}_K$. In this

**Figure 3.1.** Configuration of \mathbb{F}_{CD} vs $\tilde{\mathbb{F}}_C$.

work, we have proposed alternative -even though closely related and somewhat equivalent in the simultaneous case- definitions to generalize the notion to deflation and partial separation schemes. By contrast to Comon's terminology, the discriminacy property is here related to the local maximum points, not only the global ones. The reason is that for some criteria, we are not able, without exhaustive search, to know if a local maximum is global.

If we define the set \mathbb{F}_{CD} as the set of discriminant BSS contrasts, we know that $\mathbb{F}_{CD} \subset \mathbb{F}_C$; the discriminant contrast functions obviously forms a subset of the set of contrast functions. But, based on the previous results, the relative order of \mathbb{F}_{CD} compared to the set of contrast functions satisfying Huber's superadditivity, denoted here by $\tilde{\mathbb{F}}_C$, may be as in one of the five configurations illustrated in Figure 3.1. This section shall also help us to find the adequate configuration describing the relative order relation between the subsets \mathbb{F}_{CD} and $\tilde{\mathbb{F}}_C$ of \mathbb{F}_C .

3.2 DISCRIMINACY OF SHANNON'S ENTROPY

Before going inside technical developments, let us give three simple introductory examples in the $K = 2$ case (for illustration purposes). In these examples, the unit-norm vector \mathbf{w} can be rewritten as $\mathbf{w}_\theta \doteq [\sin \theta, \cos \theta]$ and $h(Y_\theta)$ (with $Y_\theta \doteq \mathbf{w}_\theta S$) is considered as a function of θ . The entropy is computed through Eq. (1.73), in which the pdf were estimated from a finite sample set, using Parzen density estimation with isotropic Gaussian Kernels [Parzen, 1962] and Riemannian summation instead of exact integration; this estimation of entropy will be denoted by \hat{h} (note that the value of the Parzen window is not of first importance if it is chosen in a wide reasonable range; therefore, as the curves are

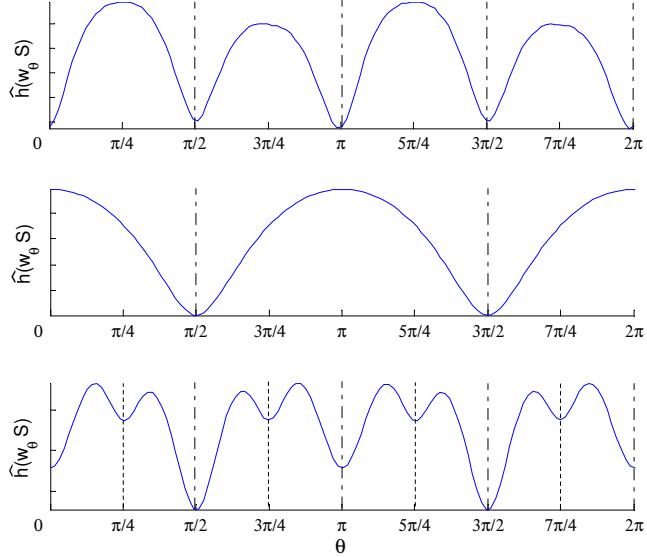


Figure 3.2. Evolution of $\hat{h}(w_\theta S)$ vs θ . Example 7: two uniform sources (top); Example 8: uniform (S_1) and Gaussian (S_2) sources (middle); Example 9: two bimodal sources (bottom). The non-mixing minima are indicated by dash-dotted vertical lines, the mixing ones by dotted lines. ©2006, IEEE. Reprinted, with permission, from Vrins, Pham & Verleysen: *Mixing and non-mixing local minima of the entropy contrast for blind source separation*. To appear in *IEEE Transactions on Information Theory* (March 2007).

given for illustration purposes only, the values of the parameters are not always provided).

Example 7 Assume that S_1 and S_2 have uniform densities. According to Theorem 11 (p. 58), local minima of the entropy exist for $\theta \in \{k\pi/2 | k \in \mathbb{Z}\}$. In this example, no mixing minimum can be observed (Fig. 3.2, top panel).

Example 8 Suppose now that S_1 and S_2 have uniform and Gaussian densities respectively. Local minima are found for $\theta \in \{(2k+1)\pi/2 | k \in \mathbb{Z}\}$, and local maxima for $\theta \in \{k\pi | k \in \mathbb{Z}\}$ (Fig. 3.2, middle panel), as expected from Theorem 11 (p. 58). Again, no spurious minimum can be observed in this example.

Example 9 Consider two “source” symmetric pdfs p_{s_1} and p_{s_2} that are constituted by i) two non-overlapping uniform modes and ii) two Gaussian modes with negligible overlap, respectively. This time, mixing maxima occur for $\theta \notin \{k\pi/2 | k \in \mathbb{Z}\}$ (Fig. 3.2, bottom panel).

In addition to an illustration of the theoretical results given by Theorem 11 (p. 58), the last example shows the existence of spurious (mixing) local minima

for $\theta \notin \{k\pi/2 | k \in \mathbb{Z}\}$. Rigorously speaking however, the figure does not constitute a proof of the existence of local minima of $h(\mathbf{wS}/\sqrt{\text{Var}[\mathbf{wS}]})$; the minima visible on the figure could indeed be a consequence of the entropy estimator (more precisely, of the pdf estimation and/or of the numerical integration as we have no idea of how $\hat{h}(Y_\theta)$ differs from the exact entropy $h(Y_\theta)$).

The phenomenon of mixing maximum point of Shannon's entropy based contrast will be investigated under several viewpoints. First, an informal approach is proposed (Section 3.2.1). Then, in Section 3.2.2, a formal approach using Taylor expansion of the entropy is provided. Finally, in Section 3.2.3 a last approach suggests to use an entropy approximation (with error bounds) to analyze the possible existence of mixing minima of Shannon's entropy. In order to avoid a lot of minus signs, this section uses the opposite of BSS contrast function, namely "cost function" (to be minimized). As an example, under some normalization constraint, Shannon's entropy is a cost function.

3.2.1 Informal approach : the multimodal case

Dealing with multimodal sources in BSS is known to be a difficult problem when achieved through a gradient descent on a cost function. Indeed, the usual cost functions plugged in the ICA algorithms may have spurious minima in such situations; the only alternative to gradient descent is the exhaustive search when no algebraic method is available [Learned-Miller and Fisher III, 2003]. As an example, consider the maximum-likelihood (ML) approach to BSS, which consists in finding an output density that is as close as possible to a target density, supposed to be – very close to – the unknown source density. The ML-based function has spurious local minima if the marginal source densities are multimodal, even if the target density is taken exactly equal to the (unknown) source density [Cardoso, 2000]. Cardoso explains intuitively that these local minima are due to a local optimal matching (in the KL divergence sense) between the output and the target densities.

This section aims at pointing out, using simulations, some specific examples where such spurious entropy minima (spurious contrast maxima) exist. The first subsection will present these examples, and a naïve reasoning is proposed to justify the location of the spurious entropy minima; these locations are shown to be related to the modality of the output densities. The second subsection aims at explaining how modality and entropy can be linked.

3.2.1.1 Structural modifications analysis: understanding the location of the entropy minima

Recently, it was noted by several authors that the entropy cost function may also have spurious minima in the BSS context [Boscolo et al., 2004, Learned-Miller and Fisher III, 2003, Vrins et al., 2004, Vrins and Verleysen, 2005b]. However, since the entropic approach does not suppose any model (there is no source model to "guess") for the source density, the existence of spurious minima

cannot be understood by the same arguments as in the ML case. These mixing maxima are emphasized in the following example.

Example 10 Consider the two pairs of sources sampled to have a scatter plot as in Figure 3.3. For both source vectors, one can compute the entropy of Y_θ

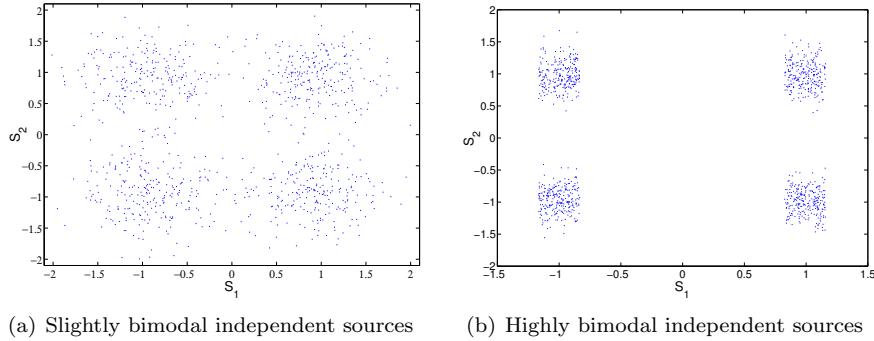


Figure 3.3. Example 10. Two scatter plots of source vectors. Both sources have two Gaussian modes (left); One source has two uniform modes, the second source has two Gaussian modes (right). ©2005, Elsevier B.V. Reprinted, with permission, from Vrins & Verleysen: *On the entropy minimization of a linear mixture of variables for source separation*. Signal Processing 85(5), pp. 1029-1044, May 2005.

(or more exactly, its approximated counterpart $\hat{h}(Y_\theta)$) as a function of θ . This is illustrated in Figure 3.4. on a polar graph. As the radius denotes the entropy, negative entropies cannot be shown; for this reason, $\hat{h}(\cdot)$ has been shifted to $\hat{h}(\cdot) + \epsilon$, where

$$\epsilon = \begin{cases} 0 & \text{if } \min_\theta \hat{h}(\cdot) \geq 0 \\ -\min_\theta \hat{h}(\cdot) & \text{if } \min_\theta \hat{h}(\cdot) < 0 \end{cases}. \quad (3.1)$$

Similarly, we can compute the sum of the output entropies. When $\mathbf{W} \in \mathcal{SO}(2)$, we can write the outputs as $Y_1 = Y_\theta$ and $Y_2 = Y_{\pi/2+\theta}$ (see mixture model in Eq. (1.60)). Figure 3.5. shows the sum $\hat{h}(Y_\theta) + \hat{h}(Y_{\pi/2+\theta})$ as a function of θ for the two pairs of sources of Figure 3.3. From these toy examples, we see that mixing minima may occur in the (sum of) entropy(ies), and that the stronger their multi-modality, the deeper the spurious minimum.

This section aims at explaining how and why these spurious minima may appear. This is done by looking at the effects of scaling and mixing independent random variables. Furthermore, this analysis allows one to understand the locations of the possible spurious minima, i.e. for which mixture coefficients they appear.

For the sake of simplicity, we still consider $K = 2$ and focus on $h(Y_\theta)$ where $Y_\theta = \mathbf{w}_\theta \mathbf{S}$ (remind that $\mathbf{w}_\theta = [\sin \theta, \cos \theta]$).

Since $\int p_{\sin \theta S_1}(\xi) d\xi = 1$, if the maximum value of $p_{\sin \theta S_1}$ increases (resp. decreases), the support $\Omega(\sin \theta S_1)$ of $\sin \theta S_1$ is contracted (resp. extended)

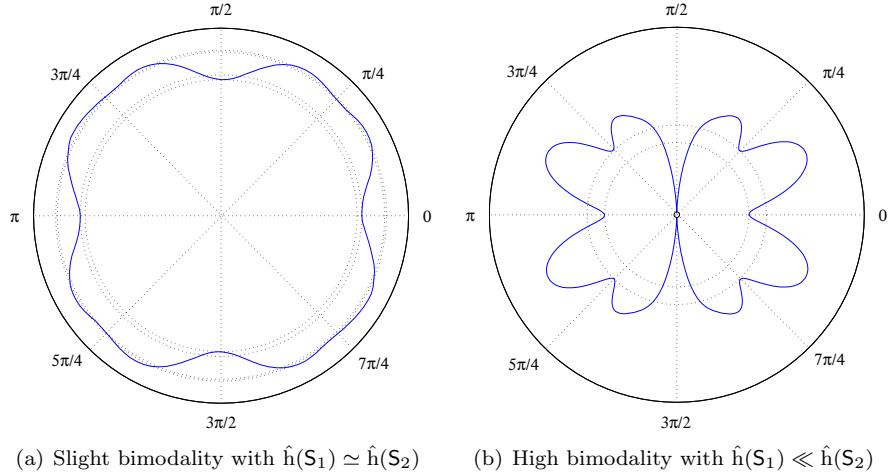


Figure 3.4. Example 10. Evolution of $\hat{h}(Y_\theta)$ vs θ for independent variables S_1, S_2 having bimodal densities (see their scatter plot in Fig. 3.3.). ©2005, Elsevier B.V. Reprinted, with permission, from Vrins & Verleysen: *On the entropy minimization of a linear mixture of variables for source separation*. *Signal Processing* 85(5), pp. 1029-1044, May 2005.

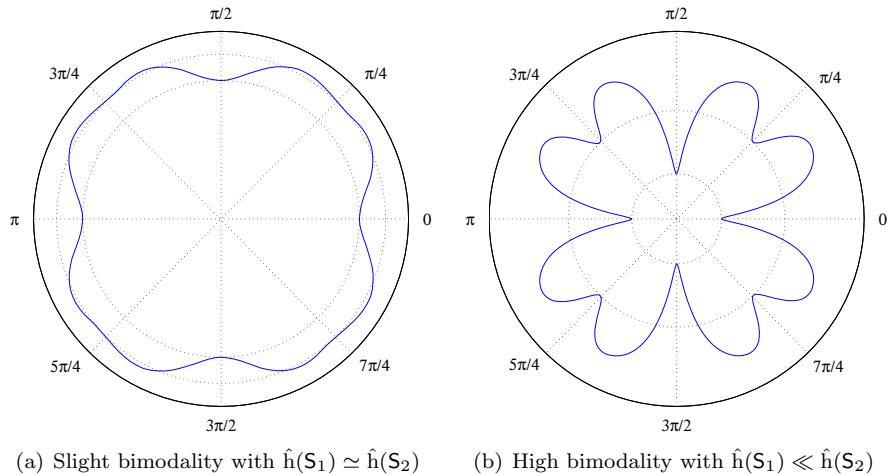


Figure 3.5. Example 10. Evolution of $\hat{h}(Y_\theta) + \hat{h}(Y_{\pi/2-\theta})$ vs θ for independent variables S_1, S_2 having bimodal densities (see their scatter plot in Fig. 3.3.). ©2005, Elsevier B.V. Reprinted, with permission, from Vrins & Verleysen: *On the entropy minimization of a linear mixture of variables for source separation*. *Signal Processing* 85(5), pp. 1029-1044, May 2005.

compared to p_{S_1} and $\Omega(S_1)$, respectively. Of course, this seems to be a nonsense if $\Omega(\sin \theta S_1)$ is infinite. Actually, this contraction/extension of the support should be understood considering ‘the inter-distances’ between the elements of $\Omega(\sin \theta S_1)$ (see below). Multiplying a variable by a real scaling coefficient smaller (resp. greater) than one contracts (resp. extends) the support of the density. Another interesting fact is that, as already mentioned in Section 2.3.2 the density p_{Y_θ} is the convolution of $p_{\sin \theta S_1}$ and $p_{\cos \theta S_2}$ [Hirschman and Widder, 1955, Feller, 1966]. Let us illustrate that point in more details in the following example.

Example 11 Consider two independent sources S_1 and S_2 with multimodal densities (see their scatter plot in Figure 3.6.(a)). We will adopt the following notation. The distance between two modes i and j of p_{S_k} will be denoted by $\Delta_{i,j}(S_k)$; it is computed between the mean of the i -th and j -th mode (“peak-to-peak”), and $i < j$. For example, we can see in Fig. 3.6. that $\Delta_{1,2}(S_1) \simeq 2$ (Fig. 3.6.(b)), $\Delta_{1,2}(S_2) \simeq 1.1$ and $\Delta_{2,3}(S_2) \simeq 1.6$ (Fig. 3.6.(c)). The entropy minima analysis is restricted to $\theta \in [0, \pi/2]$; the extension to the other quadrants of the unit circle is trivial. The solid curve on the left graph of Fig. 3.7. shows the evolution of $\hat{h}(Y_\theta)$ vs θ . The only minima relevant for source separation correspond to $\theta \in \{0, \pi/2\}$. As it can be seen on Fig. 3.7., spurious minimum appear for $\theta \neq \{0, \pi/2\}$; this is the case for several angles $\theta^\circ \simeq \{\pi/6, \pi/5, 11\pi/36\}$. These minima are thorny because in these cases, Y_θ remains a mixture of the sources; they correspond to spurious solutions.

The standard deviation, say σ_K , of the Gaussian kernels used in the pdf estimation plugged into the estimator \hat{h} may influence the quality of the estimated density. However, it seems that this is not the case (in a certain range) regarding the shape of the entropy function vs θ ; the latter is shown on the left panel of Fig. 3.7. where $\hat{h}(Y_\theta)$ is plotted vs θ for $\sigma_K \in \{0.025, 0.05, 0.1\}$. In order to improve the readability of this function, the $\sigma_K = 0.05$ curve has been plotted on a polar graph (right panel of Fig. 3.7.). To understand why such spurious minima appear at specific angles, it is useful to plot the evolution of the densities p_{Y_θ} , $p_{\sin \theta S_1}$ and $p_{\cos \theta S_2}$ vs θ . This is done in Fig. 3.8. for $\theta \in \{0, \pi/12, \pi/6, \pi/5, \pi/4, 11\pi/36, 13\pi/36, \pi/2\}$. We can observe that the critical values θ° of θ , corresponding to the spurious minima of $\hat{h}(Y_\theta)$ also minimize locally the number $N(Y_\theta)$ of modes of p_{Y_θ} . This fluctuation of $N(Y_\theta)$ as a function of the angle θ is due to the joint effect of the scaling and the mixing of the independent sources S_1 and S_2 . Obviously, $N(Y_\theta)$ is equal to $N(S_1)$ (resp. $N(S_2)$) if $\theta = \pi/2$ (resp. 0). As already explained, mixing these independent sources (keeping the variance of the mixtures unitary) has as an effect to convolute the scaled densities. Intuitively, as $N(S_1) = 2$ and $N(S_2) = 3$, when θ increases from 0 or decreases from $\pi/2$, $N(Y_\theta)$ should be equal to 6, provided that the mode width are small enough compared to the intermodal “peak-to-peak” distances. However, $N(Y_\theta)$ is not strictly increasing to a unique local maximum when θ moves apart from $k\pi/2$. The function $N(Y_\theta)$ has several local maxima for $\theta \in [0, \pi/2]$ and $\{2, 3, 6\}$ is not the whole set of acceptable values for $N(Y_\theta)$;

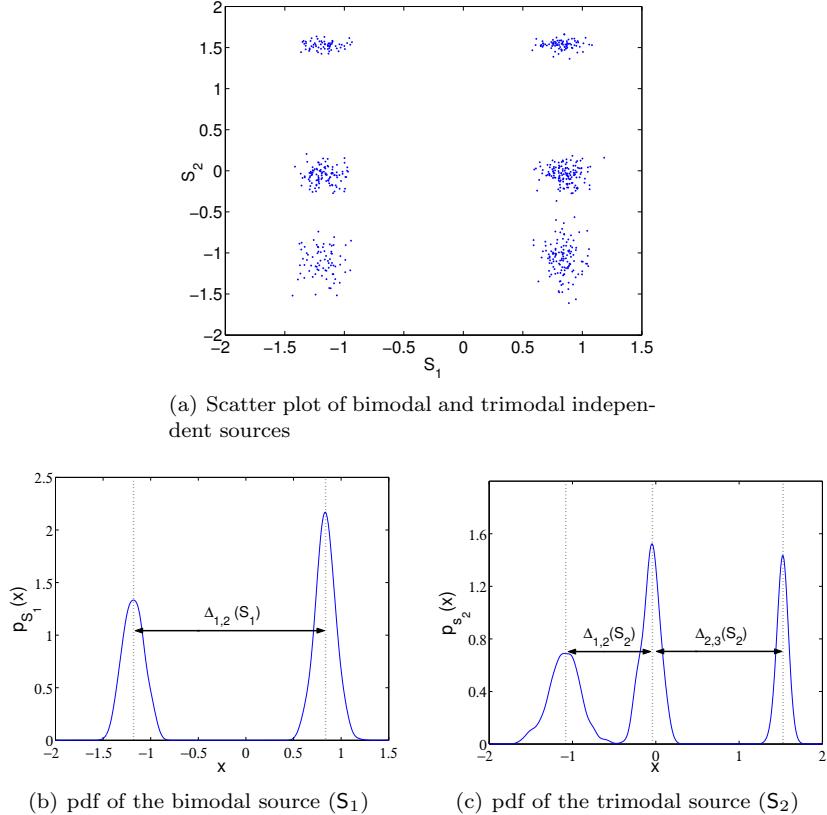


Figure 3.6. Example 11. Characteristics of the source signals S_1 and S_2 : scatter plot S_2 vs S_1 (top), p_{S_1} (bottom-left) and p_{S_2} (bottom-right). REPRINTED WITH PERMISSION FROM VRINS, ARCHAMBEAU & VERLEYSEN, BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING, AIP CONFERENCE PROCEEDINGS, VOL. 735, 589-596. © 2004, AMERICAN INSTITUTE OF PHYSICS.

p_{Y_θ} may have (locally) a particular structure if the intermodal distances of densities $p_{\sin \theta S_1}$ and $p_{\cos \theta S_2}$ become equal. In this case, $N(Y_\theta) < 6$ since two pairs of modes are superimposed during the convolution process. This situation occurs for several scaling factors of S_1 and S_2 .

As an illustration, consider the case $\theta = \pi/5$. This particular angle has the remarkable property to contract the densities p_{S_1} and p_{S_2} such that $\Delta_{1,2}(\sin(\pi/5)S_1) \simeq \Delta_{2,3}(\cos(\pi/5)S_2)$. The density of $Y_{\pi/5}$ results from the convolution of $p_{\sin(\pi/5)S_1}$ and $p_{\cos(\pi/5)S_2}$. Due to the matching of the two modes of $p_{\sin(\pi/5)S_1}$ and the two last modes of $p_{\cos(\pi/5)S_2}$ (i.e. because two pairs of modes are “simultaneously convolved”), the number of modes of Y_θ decreases: $N(Y_{\pi/5}) = 5$. The same phenomenon appears for the other values of θ° : $\Delta_{1,2}(\sin(\pi/6)S_1) \simeq \Delta_{1,2}(\cos(\pi/6)S_2)$, $\Delta_{1,2}(\sin(11\pi/6)S_1) \simeq$

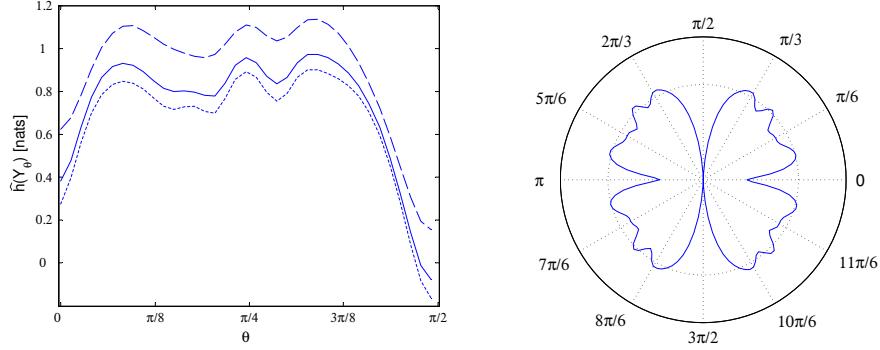


Figure 3.7. Example 11. Left: Entropy $\hat{h}(Y_\theta)$ vs θ for $\sigma_K = 0.025$ (dotted), 0.05 (solid) and 0.1 (dashed); Right: $\hat{h}(Y_\theta) + \epsilon$ (see Eq. (3.1)) vs θ ($\sigma_K = 0.05$). REPRINTED WITH PERMISSION FROM VRINS, ARCHAMBEAU & VERLEYSEN, BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING, AIP CONFERENCE PROCEEDINGS, VOL. 735, 589-596. © 2004, AMERICAN INSTITUTE OF PHYSICS.

$\Delta_{1,3}(\cos(11\pi/6)\mathsf{S}_2)$. It seems that this structural modification of $p_{Y_{\pi/5}}$ (appearing locally around θ if $\theta \in \theta^\circ$) implies a variation of the entropy.

The local mixing minima of the sum of the entropy(ies) given in Example 10 can be explained in a similar way as used in Example 11.

Note that in general, the relation that links the entropy of a variable to the number of modes of its density is not so simple: counter examples may be found easily, adjusting the width of the modes. Nevertheless, modality and entropy may be related, as sketched in the following subsection.

These results appeared in [Vrins et al., 2004, Vrins and Verleysen, 2005b].

3.2.1.2 Behind modality: understanding the existence of the entropy minima

Consider a unimodal pdf $K(y)$ of a zero-mean and unit-variance random variable. Assume that the density $p(y)$ can be written as a sum of N modes, each being a shifted and scaled version of $K(y)$, that is

$$p(y) = \sum_{n=1}^N \pi_n K_n(y) , \quad (3.2)$$

where

$$K_n(y) = \frac{1}{\sigma_n} K((y - \mu_n)/\sigma_n) \quad (3.3)$$

and π_n are positive scaling factors ensuring that $p(y)$ integrates to one (i.e. $\sum_{n=1}^N \pi_n = 1$). It is further assumed that $\mu_1 < \mu_2 < \dots < \mu_K$, without loss of generality. The support of K might be either finite or infinite, i.e. equal to \mathbb{R} .

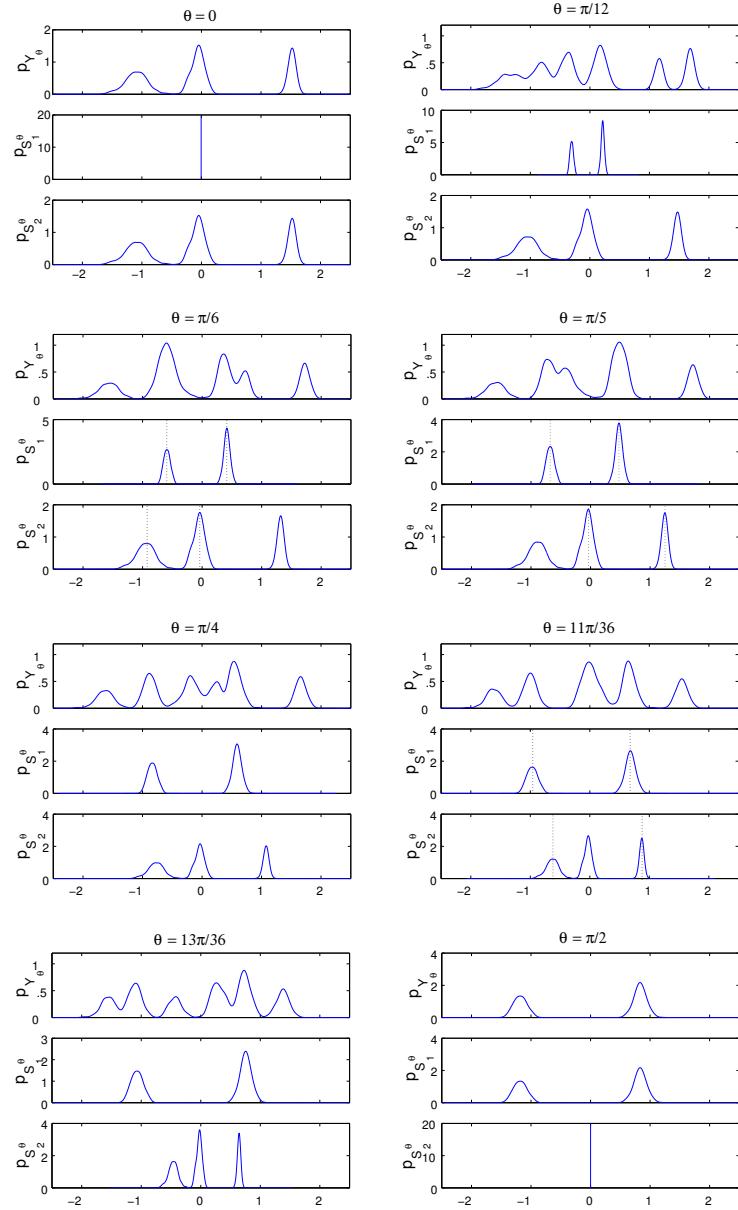


Figure 3.8. Example 11. Densities p_{Y_θ} , $p_{S_1^\theta} \doteq p_{\sin \theta S_1}$ and $p_{S_2^\theta} \doteq p_{\cos \theta S_2}$ for several values of θ . REPRINTED WITH PERMISSION FROM VRINS, ARCHAMBEAU & VERLEYSEN, BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING, AIP CONFERENCE PROCEEDINGS, VOL. 735, 589-596. © 2004, AMERICAN INSTITUTE OF PHYSICS.

However, for an unimodal pdf, a finite approximation Ω_n of $\Omega[K_n]$ is assumed to be found because the contribution of $K_n(y)$ “far from the mean of the mode” (i.e. from the μ_n) is negligible, and if the support Ω_n is centered on the mean and large enough, we have

$$\int_{\Omega_n} K_n(y) \lesssim 1 , \quad (3.4)$$

where the “ \lesssim ” symbol means “less than but close to”. We further assume that the support $\Omega[p]$ can be approximated by the union of N finite disjoint intervals Ω_n , i.e. if for all $1 \leq n \leq N$ we have

$$\int_{\Omega_n} p(y) dy \simeq \pi_n \int_{\Omega_n} K_n(y) dy \simeq \pi_n . \quad (3.5)$$

For such a strongly N -modal random variable, $h[p]$ can be approximated as follows:

$$\begin{aligned} h[p] &= - \int_{-\infty}^{+\infty} \sum_{m=1}^N \pi_m K_m(y) \log \left\{ \sum_{j=1}^N \pi_j K_j(y) \right\} dy \\ &\stackrel{(a)}{\simeq} - \sum_{n=1}^N \int_{\Omega_n} \sum_{m=1}^N \pi_m K_m(y) \log \left\{ \sum_{j=1}^N \pi_j K_j(y) \right\} dy \\ &\stackrel{(b)}{\simeq} - \sum_{n=1}^N \pi_n \int_{\Omega_n} K_n(y) \log \{ \pi_n K_n(y) \} dy \\ &= - \sum_{n=1}^N \pi_n \int_{\Omega_n} K_n(y) \left\{ \log \pi_n + \log K_n(y) \right\} dy \\ &\stackrel{(c)}{\simeq} - \sum_{n=1}^N \pi_n \left\{ \log \pi_n + \int_{\Omega_n} K_n(y) \log K_n(y) dy \right\} \\ &\stackrel{(d)}{\simeq} - \sum_{n=1}^N \pi_n \left\{ \log \pi_n + \underbrace{\int_{\mathbb{R}} K_n(y) \log K_n(y) dy}_{=-h[K_n]} \right\} \\ &= \sum_{n=1}^N \pi_n h[K_n] - \underbrace{\sum_{n=1}^N \pi_n \log \pi_n}_{\doteq -H[\boldsymbol{\pi}]} . \end{aligned} \quad (3.6)$$

Then $h[p] \approx \mathcal{H}[p]$ where

$$\mathcal{H}[p] \doteq \sum_{n=1}^N \pi_n h[K_n] + H[\boldsymbol{\pi}] . \quad (3.7)$$

In the previous development, (a) results from the multimodal form of p and $0 \log 0 = 0$ by convention. Relation (b) comes from the multi-modality of p :

in Ω_n , we can neglect the contribution of the $K_m(y)$ modes with respect to the $K_n(y)$ one, if $n \neq m$. In other words, in Ω_n , p is mainly determined by the n^{th} mode. If Ω_n is large enough, then (3.5) holds, leading to (c) and (d). Note that the approximations (a),(b),(c) and (d) are exact if $\Omega_i \cap \Omega_j = \emptyset$ for all $1 \leq i \neq j \leq N$.

In Eq. (3.7), $H[\pi]$ is the entropy of a discrete pdf with probability vector $\pi \doteq [\pi_1, \dots, \pi_n]$, and $h[K_n] = h[K] + \log \sigma_n$ (from Eq. (3.3)). In the particular case when Gaussian kernel are used, $K(y) = \phi(y)$,

$$h[K_n] = \log \sqrt{2\pi e} + \log \sigma_n . \quad (3.8)$$

When a pdf p can be efficiently modelled as a mixture of N Gaussian kernels and when approximation (3.7) holds, the pdf is said “ N -normal separable”. For such densities, the following approximator will be used for the associated entropy:

$$\mathcal{H}[p] \doteq \log \sqrt{2\pi e} + \sum_{n=1}^N \pi_n \log \sigma_n + H[\pi] . \quad (3.9)$$

The relative error $\rho[p]$ resulting from the above approximation is defined by

$$\rho[p] \doteq \left| \frac{\mathcal{H}[p] - h[p]}{h[p]} \right| . \quad (3.10)$$

Example 12 In order to illustrate the performance of this estimator on multi-normal separable variable, $h[p]$ is compared to $\mathcal{H}[p]$ on the density given in Figure 3.9. It seems reasonable to assume that p is strongly multimodal. Actually, in this example, the parameters of this 3-normal pdf are:

$$\begin{cases} \mu &= [-5, 0, 6] \\ \sigma &= [2/5, 1, 2/5] \\ \pi &= [1/6, 1/2, 1/3] \end{cases} . \quad (3.11)$$

In order to evaluate $h[p]$, the integral in the entropy definition is replaced by a Riemannian sum: $\bar{h}[p] = -\sum_x p(x) \log p(x) \Delta$, where x ranges from -10 to 10 by increasing steps of $\Delta = 5.10^{-3}$. This estimator differs from $\hat{h}[p]$ by the fact that the true analytical expression of p is used (or a sampled function expected to be very close from the true pdf, see below); it is not estimated via Parzen density estimation.

The approximation $\mathcal{H}[p]$ has been computed through (3.9). In this example, we find $\bar{\rho}[p] = 0.02\%$, where $\bar{\rho}[p]$ is a “Riemannian approximation” of $\rho[p]$ (in the sense that an exact integration has been replaced by a summation):

$$\bar{\rho}[p] \doteq \left| \frac{\mathcal{H}[p] - \bar{h}[p]}{\bar{h}[p]} \right| . \quad (3.12)$$

It is expected to be very close from $\rho[p]$ when Δ is sufficiently small, which confirms the validity of the approximator given by (3.9) for this kind of pdf.

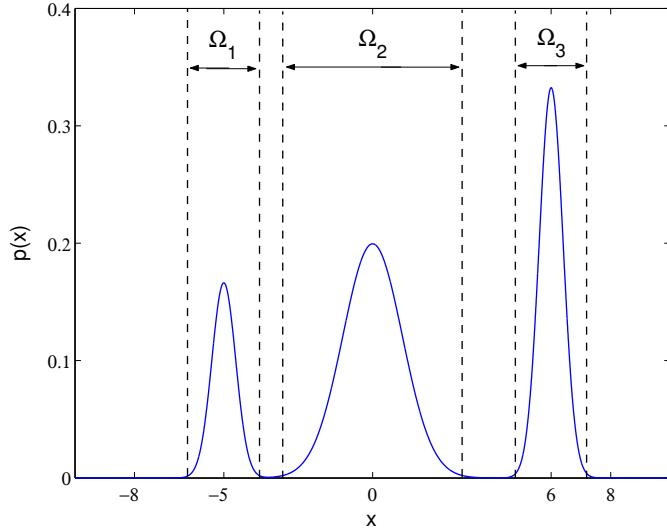


Figure 3.9. Example 12. pdf with 3 normal and approximatively disjoint modes. In this example, the Ω_i are defined as $[\mu(i) - 3\sigma(i), \mu(i) + 3\sigma(i)]$.

i	$\mu_1(S_i)$	$\mu_2(S_i)$	$\pi_1(S_i)$	$\pi_2(S_i)$	$\sigma_1(S_i) = \sigma_2(S_i)$
1	-0.995	0.995	1/2	1/2	0.1
2	-0.81	1.22	3/5	2/5	0.1

Table 3.1. Example 13. Parameters of the two bimodal pdf.

The above approximator $\mathcal{H}[p]$ can be used to indicate possible local minima of the entropy.

Example 13 Consider the bi-normal separable pdf with parameters given in Table 3.1.. They correspond to mutually independent zero-mean unit-variance source signals. The approximated entropies of these signal are: $\bar{h}(S_1) = -0.27$ and $\bar{h}(S_2) = -0.30$. Using the entropy estimator given by (3.9), we have $\max(\bar{\rho}[p_{S_1}], \bar{\rho}[p_{S_2}]) < 10^{-12}$.

As pointed out in Example 11, the number of modes of p_{Y_θ} varies with θ and, for several θ , the modes of p_{Y_θ} are Gaussian-shape with small overlap. This is due to convolution properties of Gaussian functions; the convolution of two Gaussian function is again a Gaussian function (see Appendix 2.6.8). With a slight abuse of notation, let us denote by $\mu_n(\theta), \pi_n(\theta)$ and $\sigma_n(\theta)$ the parameters of the n -th mode of the output pdf p_{Y_θ} (i.e. its mean, weight and standard deviation).

The $\mu_n(\theta), \pi_n(\theta)$ and $\sigma_n(\theta)$ parameters can be computed. For instance, the above considerations indicate that the trimodal case is obtained for $\theta_2 \doteq$

$\arctan \frac{\mu_2(S_2) - \mu_1(S_2)}{\mu_2(S_1) - \mu_1(S_1)}$ ($\pi_1(\theta_2) = 0.3$, $\pi_2(\theta_2) = 0.5$, $\pi_3(\theta_2) = 0.2$ and $\sigma_{1,2,3}(\theta_2) = 0.1$), and there also exists two angles, denoted by θ_1 and θ_3 , for which p_{Y_θ} has four approximatively disjointed modes and such that $0 < \theta_1 < \theta_2 < \theta_3 < \pi/2$. For example, we can take $\theta_3 = \frac{13\pi}{36}$ and we have $\pi_1(\theta_3) = \pi_3(\theta_3) = 0.3$, $\pi_2(\theta_3) = \pi_4(\theta_3) = 0.2$ and $\sigma_{1,2,3,4}(\theta_3) = 0.1$. On the other hand, we can choose $\theta_1 = \pi/2 - \theta_3$ ($\sigma_n(\theta_1)$ and $\pi_n(\theta_1)$ are the same as $\sigma_n(\theta_3)$ and $\pi_n(\theta_3)$, except that the values of $\pi_2(\cdot)$ and $\pi_3(\cdot)$ have to be permuted). Note that the mean of the modes do not matter as soon as the modes have a negligible overlap.

The key point is that $\mathcal{H}(Y_0 = S_2) < \mathcal{H}(Y_{\pi/2} = S_1) < \mathcal{H}(Y_{\theta_2}) < \mathcal{H}(Y_{\theta_1}) \simeq \mathcal{H}(Y_{\theta_3})$. Indeed, $\mathcal{H}(Y_{\theta_2}) = 0.21$ while $\mathcal{H}(Y_{\theta_1}) = 0.70$. In addition, p_{Y_θ} is composed of several nearly disjointed modes with Gaussian-like shapes for $\theta \in \{0, \theta_1, \theta_2, \theta_3, \pi/2\}$ and therefore, the approximator (3.9) is valid: we must have $h(Y_\theta) \simeq \mathcal{H}(Y_\theta)$ and thus a small $\rho[p_{Y_\theta}]$. As a consequence, $h(Y_\theta)$ (i.e. $\bar{h}(Y_\theta)$) must have a mixing minimum for θ in (θ_1, θ_3) . This result can be observed in Figure 3.10, where $\bar{h}(Y_\theta)$ is plotted vs θ in the first quadrant. Note that p_{Y_θ} has been computed by convoluting $p_{\sin \theta S_1}$ and $p_{\cos \theta S_2}$ through the Matlab `conv` command (this method is expected to be better than estimating the mixture pdf via a Parzen estimator for each angle). By doing so, $\bar{h}(Y_\theta)$ cannot be numerically evaluated with a high precision for $\theta \simeq k\pi/2$; this is why $\bar{h}(Y_\theta)$ is plotted for $\theta \in [\epsilon, \pi/2 - \epsilon]$, where ϵ is a small positive number. Nevertheless, it is obvious that one must have $\bar{h}(Y_\theta) \rightarrow \bar{h}(Y_{k\pi/2}) \simeq h(Y_{k\pi/2})$ when $\theta \rightarrow k\pi/2$.

The above developments are published in [Vrins et al., 2005b].

3.2.2 Formal analysis using a Taylor expansion

In this section, we shall give a proof showing the existence of spurious (mixing) entropy minima; this proof is rigorous in the sense that no approximation is used, as in the above informal approach section. The results below have been first published in [Pham, Vrins, and Verleysen, 2005, Pham and Vrins, 2005].

3.2.2.1 Simultaneous (mutual information)

To see if a point \mathbf{B} maximizes locally $\mathcal{C}_h(\mathbf{B})$ (given by Eq. (2.12)), we perform a Taylor expansion of the contrast around \mathbf{B} up to second order. Because of the multiplicative structure of the mixture model, a relative (rather than absolute) increment of the parameter \mathbf{B} is considered. More precisely, we make a Taylor development of $\mathcal{C}_h(\mathbf{B} + \mathcal{E}\mathbf{B})$ up to second order with respect to a “small matrix” \mathcal{E} (in the sense of its Frobenius norm, noted $\|\mathcal{E}\|$). Using the known result of Eq. (2.27) (see the text below the referenced equation for the definition of the notations):

$$\begin{aligned} h(Y_i + \delta Y_i) &= h(Y_i) + E[\psi_{Y_i}(Y_i)\delta Y_i] \\ &\quad + \frac{1}{2} \left\{ E[Var[\delta Y_i | Y_i]\psi'_{Y_i}(Y_i)] - (E[\delta Y_i | Y_i])'^2 \right\} \\ &\quad + o(\|\mathcal{E}\|^2) . \end{aligned} \quad (3.13)$$

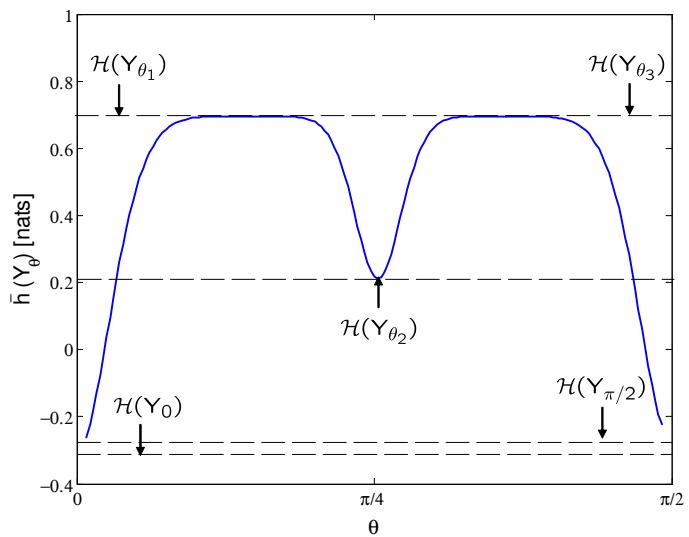


Figure 3.10. Example 13. Evolution of $\bar{h}(Y_\theta)$ for $\theta \in [\epsilon, \pi/2 - \epsilon]$, with $\epsilon \simeq 0.03$. For specific θ , p_{Y_θ} is approximatively composed of Gaussian modes with disjoint support, and for these angles, $h(Y_\theta) \simeq \mathcal{H}(Y_\theta)$. This approximation is reliable for $\theta \in \{0, [0.4, 0.55], \theta_2, [1.02, 1.17], \frac{\pi}{2}\}$.

Further, it can be shown that (see the Appendix in Section 3.8.1, p. 167)

$$\log |\det(\mathbf{B} + \mathcal{E}\mathbf{B})| = \log |\det \mathbf{B}| + \text{Tr}\mathcal{E} - \frac{1}{2}\text{Tr}\mathcal{E}^2 + o(\|\mathcal{E}\|^2) . \quad (3.14)$$

Therefore, with $\delta\mathbf{Y}_i = \sum_{k=1}^K \mathcal{E}_{ik}\mathbf{Y}_k$, noting that $E[\psi_{\mathbf{Y}_k}(\mathbf{Y}_k)\mathbf{Y}_k] = 1$ (see Property 4 and [Pham, 1996]), we have

$$\begin{aligned} \mathcal{C}_h(\mathbf{B} + \mathcal{E}\mathbf{B}) &= \mathcal{C}_h(\mathbf{B}) - \sum_{i \neq j} \sum E[\psi_{\mathbf{Y}_i}(\mathbf{Y}_i)\mathbf{Y}_j]\mathcal{E}_{ij} - \sum_{i=1}^K \sum_{j=1}^K \sum_{k=1}^K \frac{\mathcal{E}_{ij}\mathcal{E}_{ik}}{2} \times \\ &\quad \left\{ E[\text{Cov}(\mathbf{Y}_j, \mathbf{Y}_k | \mathbf{Y}_i)\psi'_{\mathbf{Y}_i}(\mathbf{Y}_i)] + (E[\mathbf{Y}_j | \mathbf{Y}_i])'(E[\mathbf{Y}_k | \mathbf{Y}_i])' \right\} \\ &\quad - \frac{1}{2} \sum_{i,j} \mathcal{E}_{ij}\mathcal{E}_{ji} + o(\|\mathcal{E}\|^2) , \end{aligned} \quad (3.15)$$

where $\text{Cov}(\mathbf{Y}_j, \mathbf{Y}_k | \mathbf{Y}_i) = E[\mathbf{Y}_j\mathbf{Y}_k | \mathbf{Y}_i] - E[\mathbf{Y}_j | \mathbf{Y}_i]E[\mathbf{Y}_k | \mathbf{Y}_i]$.

The above expansion shows that \mathbf{B} is a stationary point of $\mathcal{C}_h(\mathbf{B})$ if $E[\psi_{\mathbf{Y}_i}(\mathbf{Y}_i)\mathbf{Y}_j] = 0$ for $i \neq j$. To see if \mathbf{B} is indeed a local maximum, one has to look at the second order term in the above expansion, which is quite involved. Therefore, we shall focus on the case of two sources ($K = 2$).

Consider the $K = 2$ case where both sources share the same density function p_S . Since $\text{KL}([\mathbf{Y}_1, \mathbf{Y}_2]) \geq 0$ with equality if and only if the variables $\mathbf{Y}_1, \mathbf{Y}_2$ are mutually independent, $\mathcal{C}_h(\mathbf{B})$ admits a global maximum when $\mathbf{B} \sim \mathbf{A}^{-1}$ as for these demixing matrices, the outputs are proportional to distinct sources. We now show that for a certain source density p_S , the point

$$\mathbf{B}^\diamond \doteq \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \mathbf{A}^{-1} \quad (3.16)$$

is also a local maximum of $\mathcal{C}_h(\mathbf{B})$ whatever is \mathbf{A} in $\mathcal{M}(2)$. This remains true if the above right hand side is left multiplied by $\mathbf{P}\mathbf{D}$, since $\mathcal{C}_h(\mathbf{B})$ is invariant when one left multiplies its argument by any permutation or diagonal matrix.

For \mathbf{B}^\diamond given in (3.16), $\mathbf{Y}_1 = \mathbf{S}_1 + \mathbf{S}_2$, $\mathbf{Y}_2 = \mathbf{S}_2 - \mathbf{S}_1$. Therefore

$$E[\psi_{\mathbf{Y}_1}(\mathbf{Y}_1)\mathbf{Y}_2] = E[\psi_{\mathbf{Y}_1}(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{S}_2] - E[\psi_{\mathbf{Y}_1}(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{S}_1] , \quad (3.17)$$

$$E[\psi_{\mathbf{Y}_2}(\mathbf{Y}_2)\mathbf{Y}_1] = E[\psi_{\mathbf{Y}_2}(\mathbf{S}_2 - \mathbf{S}_1)\mathbf{S}_1] + E[\psi_{\mathbf{Y}_2}(\mathbf{S}_2 - \mathbf{S}_1)\mathbf{S}_2] . \quad (3.18)$$

But since the joint densities of $(\mathbf{S}_1, \mathbf{S}_2)$ and of $(\mathbf{S}_2, \mathbf{S}_1)$ are the same, one can permute \mathbf{S}_1 and \mathbf{S}_2 in the above right hand sides without affecting their value. Hence these right hand sides vanish, noting that \mathbf{Y}_2 has the same density as $-\mathbf{Y}_2$ and hence $\psi_{\mathbf{Y}_2}$ is an odd function.

The above results show that \mathbf{B}^\diamond is a stationary point of $\mathcal{C}_h(\mathbf{B})$. To see if it is a local maximum point, we consider the expansion of $\mathcal{C}_h(\mathbf{B}^\diamond + \mathcal{E}\mathbf{B}^\diamond)$ up to second order. Again, since one can permute \mathbf{S}_1 and \mathbf{S}_2 without changing their joint densities, $E[\mathbf{S}_2|\mathbf{S}_1 + \mathbf{S}_2] = E[\mathbf{S}_1|\mathbf{S}_2 + \mathbf{S}_1]$ and hence $E[\mathbf{Y}_2|\mathbf{Y}_1] = 0$. In the case where p_S is symmetric so that $-\mathbf{S}_1$ has the same density as \mathbf{S}_1 , by the same

argument as before with S_1 replaced by $-S_1$: $E[S_2|S_2 - S_1] = -E[S_1|S_2 - S_1]$ and hence $E[Y_1|Y_2] = 0$. Therefore from the result of Eq. (2.27) and noting that [Gray and Davisson, 2004]

$$E[E[f(X, Y)|X]g(X)] = E[f(X, Y)g(X)] , \quad (3.19)$$

we have $E[E[Y_j^2|Y_i]\psi'_{Y_i}(Y_i)] = E[Y_j^2\psi'_{Y_i}(Y_i)]$ and :

$$\begin{aligned} \mathcal{C}_h(\mathbf{B}^\diamond + \mathcal{E}\mathbf{B}^\diamond) &= \mathcal{C}_h(\mathbf{B}^\diamond) - \frac{1}{2} \left\{ E[Y_2^2\psi'_{Y_1}(Y_1)]\mathcal{E}_{12}^2 + E[Y_1^2\psi'_{Y_2}(Y_2)]\mathcal{E}_{21}^2 \right\} \\ &\quad - \mathcal{E}_{12}\mathcal{E}_{21} + o(\|\mathcal{E}\|^2) . \end{aligned} \quad (3.20)$$

Observe that for $i \neq j$ the iterative expectation lemma summarized in Eq. (3.19) gives (with $X = Y_i$, $Y = Y_j$, $f(X, Y) = XY$ and $g(X) = 1$) $E[E[Y_j Y_i|Y_i]] = E[Y_j Y_i]$. The last expectation vanishes by noting the relation between Y_i , Y_j and the sources.

But the sum of the last two term equals

$$-\frac{1}{2} \{ E[Y_2^2\psi'_{Y_1}(Y_1)]\mathcal{E}_{12}^2 + E[Y_1^2\psi'_{Y_2}(Y_2)]\mathcal{E}_{21}^2 + 2\mathcal{E}_{12}\mathcal{E}_{21} \} ,$$

and the term into braces can be written as the following quadratic form:

$$[\mathcal{E}_{12} \quad \mathcal{E}_{21}] \begin{bmatrix} E[Y_2^2\psi'_{Y_1}(Y_1)] & 2 \\ 0 & E[Y_1^2\psi'_{Y_2}(Y_2)] \end{bmatrix} [\mathcal{E}_{12} \quad \mathcal{E}_{21}]^T \quad (3.21)$$

which is positive definite if and only if $4 < 4E[Y_2^2\psi'_{Y_1}(Y_1)]E[Y_1^2\psi'_{Y_2}(Y_2)]$ [Brookes, 2005]. The above expansion shows that \mathbf{B}^\diamond given in (3.16) is a local maximum point of $\mathcal{C}_h(\mathbf{B})$ if and only if

$$E[Y_2^2\psi'_{Y_1}(Y_1)]E[Y_1^2\psi'_{Y_2}(Y_2)] > 1 . \quad (3.22)$$

But since the joint density of (Y_1, Y_2) is the same as the one of (Y_2, Y_1) , this condition is equivalent to

$$E[Y_2^2\psi'_{Y_1}(Y_1)] > 1 . \quad (3.23)$$

The following example shows that such a mixing maximum may exist in a specific and realistic situation, that is for a simple source density p_S .

Example 14 *Inspired by simulations given in the informal approach section, we now show that the above condition (3.23) is satisfied (so that \mathbf{B} given in (3.16) is a spurious local maximum point) for a particular source density being a mixture of densities of the form $\sum_{n=1}^N \frac{\pi_n}{\sigma_n} K\left(\frac{y - \mu_n}{\sigma_n}\right)$ if $N = 2$, $K(\cdot) = \phi(\cdot)$, $\boldsymbol{\pi} = [0.5, 0.5]$, $\mu = [-1, 1]$ and $\sigma_n = [\sigma, \sigma]$, where $\phi(s)$ is the standard normal density defined in Eq. (2.56). This density is precisely*

$$p_S(s) = \{\phi[(s+1)/\sigma] + \phi[(s-1)/\sigma]\}/(2\sigma) , \quad (3.24)$$

and we shall show that Eq. (3.23) holds for sufficiently small σ (i.e. when p_S is “bimodal enough”): a local mixing maximum of $C_h(\mathbf{B})$ exists at $\mathbf{B} = \mathbf{B}^\diamond$. In this specific example, the following two lemmas can be proved (the proofs are relegated to Section 3.8.2 p. 169 and Section 3.8.3 p. 170, respectively).

Lemma 11 Let S_1 and S_2 have the same density p_S . Then $Y_1 = S_1 + S_2$ and $Y_2 = S_2 - S_1$ also have the same density

$$p_Y(y) = \frac{1}{4\sqrt{2}\sigma} \phi\left(\frac{y+2}{\sqrt{2}\sigma}\right) + \frac{1}{2\sqrt{2}\sigma} \phi\left(\frac{y}{\sqrt{2}\sigma}\right) + \frac{1}{4\sqrt{2}\sigma} \phi\left(\frac{y-2}{\sqrt{2}\sigma}\right). \quad (3.25)$$

Their common score function ψ_Y admits the derivative

$$\psi'_Y(y) = \frac{1}{2\sigma^2} - \frac{w_{-1}(y)w_0(y) + w_1(y)w_0(y) + 4w_{-1}(y)w_1(y)}{\sigma^4}, \quad (3.26)$$

where

$$w_0(y) = \frac{2\phi(y/\sqrt{2}\sigma)}{\phi[(y+2)/\sqrt{2}\sigma] + 2\phi(y/\sqrt{2}\sigma) + \phi[(y-2)/\sqrt{2}\sigma]},$$

$$w_{\mp 1}(y) = \frac{\phi[(y \pm 2)/\sqrt{2}\sigma]}{\phi[(y+2)/\sqrt{2}\sigma] + 2\phi(y/\sqrt{2}\sigma) + \phi[(y-2)/\sqrt{2}\sigma]}.$$

Further, $E[Y_2^2|Y_1 = y] = 2\sigma^2 + 4w_0(y)$.

Lemma 12 The expectation $E[Y_2^2\psi'_{Y_1}(Y_1)]$ equals

$$1 + \frac{1}{\sigma^2} - \int \frac{[\sigma^2 + 2w_0(y)][w_0(y) + 2w_1(y)]}{\sigma^4} \phi\left(\frac{y+2}{\sqrt{2}\sigma}\right) dy. \quad (3.27)$$

The last term in the above expression tends to 0 as $\sigma \rightarrow 0$ and hence $E[Y_2^2\psi'_{Y_1}(Y_1)] \rightarrow \infty$ as $\sigma \rightarrow 0$.

The above results show that for σ small enough, there is a spurious minimum at the point $\mathbf{B} = \mathbf{B}^\diamond$ (3.16) as the left-hand side member of Eq. (3.23) tends to infinity as $\sigma \rightarrow 0$, by Lemma 12. Figure 3.11. illustrates the case $\sigma = 0.7$ for which $E[Y_2^2\psi'_{Y_1}(Y_1)] = 0.9489 < 1$ and Figure 3.12. and Figure 3.13. illustrate the cases $\sigma = 0.5$ and $\sigma = 0.2$ for which $E[Y_2^2\psi'_{Y_1}(Y_1)] = 1.5412 > 1$ and $E[Y_2^2\psi'_{Y_1}(Y_1)] = 25.71 > 1$, respectively. One can see from figures Fig. 3.11.(a), Fig. 3.12.(a) and Fig. 3.13.(a) that as σ decreases, p_Y changes from a unimodal to a trimodal structure, and from figures Fig. 3.11.(b), Fig. 3.12.(b) and Fig. 3.13.(b) that $2\sigma^2\psi'_Y$ approaches 1 inside the three regions $(-\infty, -1), (-1, 1), (1, \infty)$, with “dips” at the transition points (the “dips” being sharper for smaller σ). The product of ψ'_Y with the function $y \mapsto E[Y_2^2|Y_1 = y]$, which equals $2\sigma^2 + 4w_0$, produces a curve of similar shape as $2\sigma^2\psi'_Y$, but with a higher level in the central region due to the term $4w_0$. Its integral with respect to the density p_Y yields

$$\int p_Y(y)E[Y_2^2|Y_1 = y]\psi'_Y(y)dy \stackrel{(3.19)}{=} E[Y_2^2\psi'_{Y_1}(Y_1)]. \quad (3.28)$$

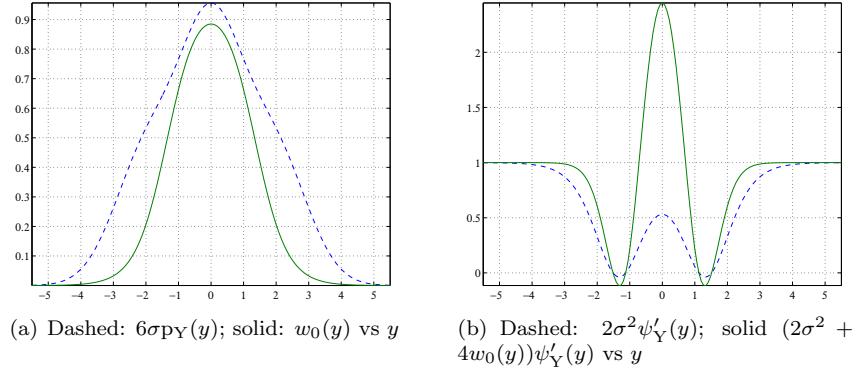


Figure 3.11. Example 14. Density $p_Y(y)$ is unimodal ($\sigma = 0.7$, $E[Y_2^2 \psi'_{Y_1}(Y_1)] \simeq .95$): spurious minima of mutual information and entropy cannot be observed (see text).

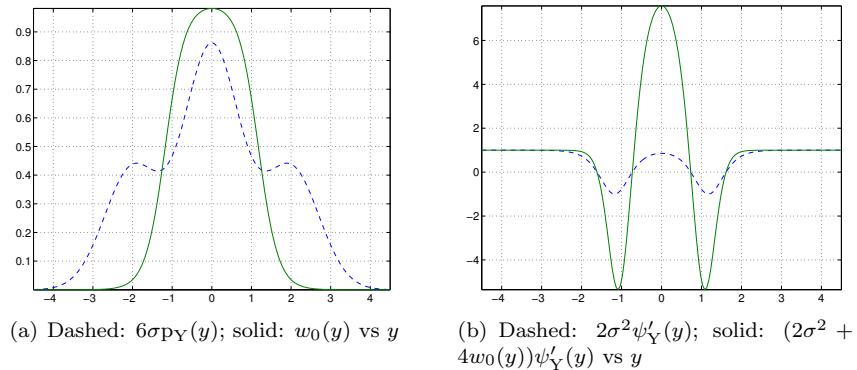


Figure 3.12. Example 14. Density $p_Y(y)$ is slightly trimodal ($\sigma = 0.5$, $E[Y_2^2 \psi'_{Y_1}(Y_1)] \simeq 1.54$): spurious minima of mutual information and entropy can be observed (see text).

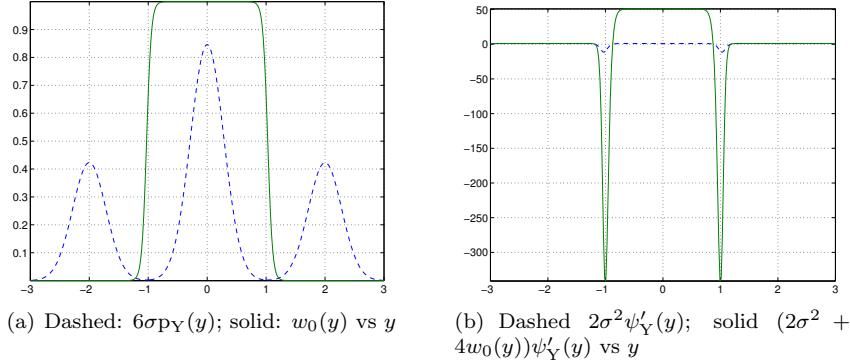


Figure 3.13. Example 14. Density $p_Y(y)$ is strongly trimodal ($\sigma = 0.2$, $E[Y_2^2 \psi'_{Y_1}(Y_1)] \simeq 25.71$): deep spurious minima of mutual information and entropy can be observed (see text).

As p_Y is low in the neighborhood of the transition points ± 1 , even more so as σ becomes smaller, the effect of the “dips” is attenuated and the integral should become larger as σ decreases, since the function takes a higher value inside the central region. This explains why one gets a larger value of $E[Y_2^2 \psi'_{Y_1}(Y_1)]$. Figure 3.14. plots this quantity versus σ . One can see that when σ decreases beyond the value 0.63 (approximately) this quantity becomes greater than 1.

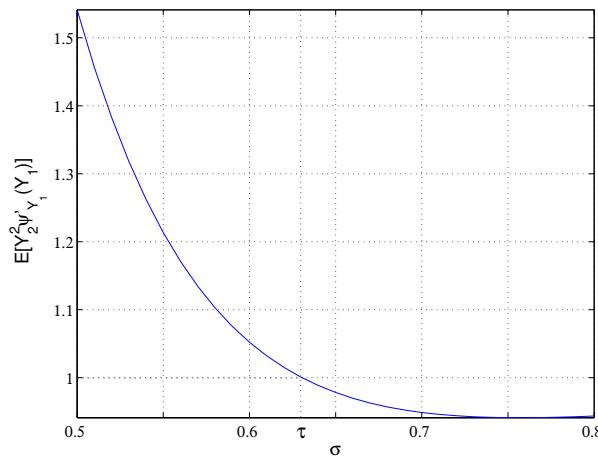


Figure 3.14. Example 14. Plot of $E[Y_2^2 \psi'_{Y_1}(Y_1)]$ vs σ . ©2005, IEEE. Reprinted, with permission, from Pham & Vrins: Local Minima of Information-Theoretic Criteria in Blind Source Separation. IEEE Signal Processing Letters 12(11), pp. 788-791, November 2005.

3.2.2.2 Deflation (negentropy)

We consider the negentropy based FastICA, which consists in maximizing the negentropy of $\mathbf{Y}_i = \mathbf{b}_i \mathbf{X}$ (see Section 2.2.2.1). One may assume that the \mathbf{S}_k have the same variance, since one can divide any of them by a constant and multiply the corresponding columns of \mathbf{A} by the same constant, without changing the mixture model. Since the negentropy is scale invariant, one may assume $\text{Var}[\mathbf{Y}_i] = 1$ (see also assumption \mathcal{A}_7). Thus, negentropy based FastICA amounts to minimizing $h(\mathbf{w}_i \mathbf{S})$ under the constraint $\|\mathbf{w}_i\| = 1$.

Again, considering the case of two sources and let $\mathbf{Y}_\theta = \sin \theta \mathbf{S}_1 + \cos \theta \mathbf{S}_2 = \mathbf{w}_\theta \mathbf{S}$ first defined in Eq. (1.61), it is easy to see that a small change $\delta\theta$ in θ induces a change

$$\delta \mathbf{Y}_\theta = \underbrace{(\cos \theta \mathbf{S}_1 - \sin \theta \mathbf{S}_2)}_{-\mathbf{Y}_{\theta+\pi/2}} \delta\theta - \frac{1}{2} \mathbf{Y}_\theta \delta\theta^2 + o(\delta\theta^2) \quad (3.29)$$

in \mathbf{Y}_θ up to second order in $\delta\theta$. Thus by the same calculation as in Section 2.2.2.2, using Eq. (2.27), one gets, putting $\mathbf{Y}_\theta^\perp = -\mathbf{Y}_{\theta+\pi/2}$ and noting that $E[\psi_{\mathbf{Y}_\theta}(\mathbf{Y}_\theta)\mathbf{Y}_\theta] = 1$ (see Property 4, p. 57):

$$\begin{aligned} h(\mathbf{Y}_\theta + \delta \mathbf{Y}_\theta) &\approx h(\mathbf{Y}_\theta) + \delta\theta E[\psi_{\mathbf{Y}_\theta}(\mathbf{Y}_\theta)\mathbf{Y}_\theta^\perp] \\ &\quad + \frac{\delta\theta^2}{2} \{E[\text{Var}[\mathbf{Y}_\theta^\perp|\mathbf{Y}_\theta]\psi'_{\mathbf{Y}_\theta}(\mathbf{Y}_\theta)] - (E[\mathbf{Y}_\theta^\perp|\mathbf{Y}_\theta])'^2 - 1\}, \end{aligned} \quad (3.30)$$

up to second order in $\delta\theta$.

The above result shows that a stationary point of $h(\mathbf{Y}_\theta)$ (as a function of θ) occurs when $E[\psi_{\mathbf{Y}_\theta}(\mathbf{Y}_\theta)\mathbf{Y}_\theta^\perp] = 0$. Clearly, this is achieved for $\theta = 0$ and $\theta = \pi/2$ since $\mathbf{Y}_0 = -\mathbf{Y}_{\pi/2}^\perp = \mathbf{S}_2$, $\mathbf{Y}_{\pi/2} = \mathbf{Y}_0^\perp = \mathbf{S}_1$ and \mathbf{S}_1 and \mathbf{S}_2 are independent. The points $\theta = 0$ and $\theta = \pi/2$ are actually local minima of $h(\mathbf{Y}_\theta)$ if \mathbf{p}_S is non Gaussian. Indeed, the second derivative of $h(\mathbf{Y}_\theta)$ at $\theta = 0$ and $\theta = \pi/2$ reduces to $\text{Var}[\mathbf{S}_1]E[\psi'_{\mathbf{S}_2}(\mathbf{S}_2)] - 1$ and $\text{Var}[\mathbf{S}_2]E[\psi'_{\mathbf{S}_1}(\mathbf{S}_1)] - 1$ respectively. But for any random variable \mathbf{Y} , $E[\psi'_Y(Y)] = E[\psi_Y^2(Y)]$ by integration by parts and $\text{Var}[\mathbf{Y}]E[\psi_Y^2(\mathbf{Y})] \geq 1$ by the Schwartz inequality (noting that $E[\psi_Y(\mathbf{Y})\mathbf{Y}] = 1$ and $E[\psi_Y(\mathbf{Y})] = 0$, see Property 4, p. 57). The inequality is strict unless ψ_Y is linear, that is \mathbf{Y} is Gaussian. Note that since $h(\mathbf{Y}_\theta)$ is periodic (with respect to θ) of period π , the function $h(\mathbf{Y}_\theta)$ admits local minima for θ in $\{p\pi/2 \mid p \in \mathbb{Z}\}$.

The same arguments as in Section 3.2.2 show that there are two other stationary points of $h(\mathbf{Y}_\theta)$ at $\theta = \pi/4$, for which $\mathbf{Y}_\theta = \mathbf{Y}_1/\sqrt{2}$ and $\mathbf{Y}_\theta^\perp = \mathbf{Y}_2/\sqrt{2}$, and at $\theta = 3\pi/4$, for which $\mathbf{Y}_\theta = -\mathbf{Y}_2/\sqrt{2}$ and $\mathbf{Y}_\theta^\perp = -\mathbf{Y}_1/\sqrt{2}$. To see if they are the local minima, we look at the second derivative of $h(\mathbf{Y}_\theta)$. As before, $E[\mathbf{Y}_2|\mathbf{Y}_1] = 0$ and if \mathbf{p}_S is symmetric, $E[\mathbf{Y}_1|\mathbf{Y}_2] = 0$. In this case, the second derivative of $h(\mathbf{Y}_\theta)$ at $\theta = \pi/4$ and $\theta = 3\pi/4$ reduces to $E[\mathbf{Y}_2^2\psi_{\mathbf{Y}_1}(\mathbf{Y}_1)] - 1$ and $E[\mathbf{Y}_1^2\psi_{\mathbf{Y}_2}(\mathbf{Y}_2)] - 1$, respectively, which are equal. Thus the condition for these points are local maxima of the negentropy are the same as the condition for the point (3.16) to be a local minimum of the mutual information criterion (given by Eq. (3.23)).

Figure 3.15. shows the approximated log-entropies of \mathbf{Y}_θ (i.e. $\log h(\mathbf{Y}_\theta)$) versus the mixing angle θ for some values of σ . When σ decreases beyond 0.63

(approximately, see Fig. 3.14.) this quantity becomes greater than 1 and thus local minima can be observed in Fig. 3.15. (even though it is difficult for $\sigma = 0.5$).

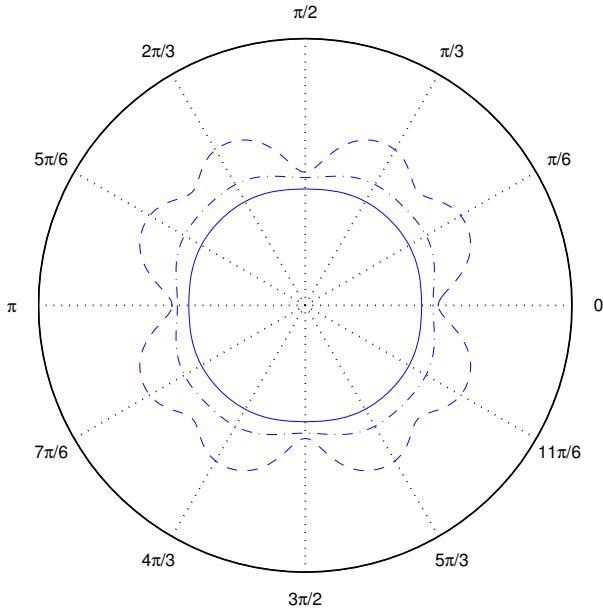


Figure 3.15. Effect of σ on the entropy mixing minima for the sources of the example: $\log \bar{h}(Y_\theta)$ vs θ for $\sigma = 0.7$ (solid), $\sigma = 0.5$ (dash-dot), $\sigma = 0.2$ (dashed).

3.2.3 Formal analysis using entropy approximation

Entropy bounds will be used in the next section to prove that for a specific class of source densities, the entropy function $h(\mathbf{wS})$ can have a local minimum that does not correspond to a row of the identity matrix. The presented approach yields more general results than those of Section 3.2.2, since it is no longer required that the sources share a common symmetric pdf.

This approach relies on an entropy approximation of a multimodal pdf of the form given by Eq. (3.2) where $N > 1$, π_1, \dots, π_N are (strictly positive) probabilities summing to 1 and K_1, \dots, K_N are unimodal pdfs. We focus on the case where the supports of the K_n can be nearly covered by disjoint subsets Ω_n ($n = 1, \dots, N$) so that p is strongly multimodal (with N modes). In this case a good approximation to the entropy of a random variable of density p can be obtained; this entropy will be *abusively* denoted by $h[p]$ instead of $h(Y)$ where Y is a random variable with pdf p . Such an approximation will first be derived informally (for ease of comprehension) and then a formal development giving the

error bounds of the approximator is provided. This work has been addressed in [Pham, Vrins, and Verleysen, 2005, Vrins, Pham, and Verleysen, 2007b].

3.2.3.1 Upper and lower bounds on the entropy of a multimodal density

The entropy approximator $\mathcal{H}[p]$ given by Eq. (3.7) is actually an upper bound for the entropy. This claim is proved in what follows; in addition, a lower bound of the entropy will be further provided. These bounds permit us to analyze how accurate is the approximation $h[p] \approx \mathcal{H}[p]$ (in a worst-case sense); they are explicitly computed when all K_n are Gaussian kernels.

The following lemma provides upper and lower bounds for the entropy.

Lemma 13 *Let p be given by Eq. (3.2), then*

$$h[p] \leq \mathcal{H}[p] , \quad (3.31)$$

where $\mathcal{H}[p]$ is given by Eq. (3.7).

In addition, assume that $\sup K_n = \sup_{y \in \mathbb{R}} K_n(y) < \infty$ ($1 \leq n \leq N$) and let $\Omega_1, \dots, \Omega_N$ be disjoint subsets which approximately cover the supports of K_1, \dots, K_N , in the sense that

$$\begin{cases} \epsilon_n \doteq \int_{\mathbb{R} \setminus \Omega_n} K_n(y) dy , \\ \epsilon'_n \doteq \int_{\mathbb{R} \setminus \Omega_n} K_n(y) \log \frac{\sup K_n}{K_n(y)} dy \end{cases}$$

are small. Then, we have

$$h[p] \geq \mathcal{H}[p] - \left(\underbrace{\sum_{n=1}^N \pi_n \epsilon'_n + \sum_{n=1}^N \pi_n \left[\log \left(\frac{\max_{1 \leq m \leq N} \sup K_m}{\pi_n \sup K_n} \right) + 1 \right] \epsilon_n}_{\doteq \Xi} \right) \quad (3.32)$$

where $\Xi \geq 0$.

The proof of this lemma is given in the Appendix of the Chapter, in Section 3.8.4, p. 171.

Let us consider now the case where the densities K_n in (3.2) all have the same form:

$$K_n(y) = (1/\sigma_n) K[(y - \mu_n)/\sigma_n] , \quad (3.33)$$

where K is a bounded density of finite entropy. Hence $h[K_n] = h[K] + \log \sigma_n$ and the upper bound (3.7) becomes

$$\mathcal{H}[p] = h[K] + \sum_{n=1}^N \pi_n \log \sigma_n + H(\boldsymbol{\pi}) . \quad (3.34)$$

Also, the lower bound on the entropy given by Eq. (3.32) reduces to

$$\mathcal{H}[p] - \underbrace{\sum_{n=1}^N \pi_n [\epsilon'_n + (\log \pi_n^{-1} + 1)\epsilon_n]}_{\Xi}. \quad (3.35)$$

Let us arrange the μ_n by increasing order and take σ_n small with respect to

$$d_n \doteq \min(\mu_n - \mu_{n-1}, \mu_{n+1} - \mu_n), \quad (3.36)$$

where $\mu_0 = -\infty$ and $\mu_{N+1} = \infty$ by convention. Under this assumption, the density (3.2) is strongly multimodal and the Ω_n in the above lemma can be taken to be intervals centered at μ_n of length d_n :

$$\Omega_n \doteq (\mu_n - d_n/2, \mu_n + d_n/2). \quad (3.37)$$

Then simple calculations give

$$\begin{cases} \epsilon_n &= 1 - \int_{-\mu_n/(2\sigma_n)}^{d_n/(2\sigma_n)} K(x)dx, \\ \epsilon'_n &= h[K] - H_{d_n/\sigma_n}(K) + \epsilon_n \log(\sup K), \end{cases}$$

where $H_\alpha(K) \doteq -\int_{-\alpha/2}^{\alpha/2} K(x) \log K(x)dx$. It is clear that ϵ_n and ϵ'_n both tend to 0 as $d_n/\sigma_n \rightarrow \infty$. Thus one gets the following corollary.

Corollary 12 *Let p be given by (3.2) with K_n of the form (3.33) and $\sup_x K(x) < \infty$. Then $h[p]$ is bounded above by $\mathcal{H}[p]$ and converges to this bound as $\min_n(d_n/\sigma_n) \rightarrow \infty$, d_n being defined in (3.36).*

Let us now focus on the $K(x) = \phi(x)$ case, which means that we restrict our analysis to densities that are “mixture of Gaussian functions”.

The upper and lower bounds on $h[p]$ are given by (3.34) and (3.35) with $h[\phi]$ instead of $h[K]$; ϵ_n and ϵ'_n can now be obtained explicitly :

$$\begin{cases} \epsilon_n = \text{Erfc}\left(\frac{d_n}{2\sqrt{2}\sigma_n}\right), \\ \epsilon'_n = h[\phi] - H_{d_n/\sigma_n}(\phi) - \epsilon_n \log \sqrt{2\pi}, \end{cases}$$

where Erfc is the complementary error function defined as $\text{Erfc}(x) = (2/\sqrt{\pi}) \int_x^\infty e(-z^2) dz$. By double integration by parts and noting that $\int \text{Erf}(x) dx = x\text{Erf}(x) + e(-x^2)/\sqrt{\pi}$ with $\text{Erf}(x) = 1 - \text{Erfc}(x)$, some algebraic manipulations give

$$H_{d_n/\sigma_n}(\phi) = \frac{1}{2} \text{Erf}\left(\frac{d_n}{2\sqrt{2}\sigma_n}\right) \log(2\pi e) - \frac{d_n}{2\sqrt{2\pi}\sigma_n} e^{-d_n^2/(8\sigma_n^2)}.$$

One can see that $H_{d_n/\sigma_n}(\phi) \rightarrow h[\phi] = \log \sqrt{2\pi e}$ as $d_n/\sigma_n \rightarrow \infty$, as it should be. Finally:

$$\begin{cases} \epsilon_n &= \operatorname{Erfc}\left(\frac{d_n}{2\sqrt{2}\sigma_n}\right) \\ \epsilon'_n &= \frac{1}{2}\operatorname{Erfc}\left(\frac{d_n}{2\sqrt{2}\sigma_n}\right) + \frac{d_n}{2\sqrt{2\pi}\sigma_n}e^{-[d_n/(2\sqrt{2}\sigma_n)]^2} \end{cases}.$$

Example 15 To illustrate Corollary 12, Fig. 3.16. plots the entropy of a trimodal variable Y with density p as in (3.2), K_n given by (3.33), $\sigma_n = \sigma$ (for the ease of illustration), $K = \phi$, $\mu = [0, 5, 10]$ and $\pi = [1/4, 1/2, 1/4]$. Such a variable can be represented as $Y = U + \sigma Z$ where U is a discrete random variable taking values in $\{0, 5, 10\}$ with probabilities $1/4, 1/2, 1/4$ respectively and Z is a standard Gaussian variable independent from U . The upper and lower bounds on the entropy are computed as in Lemma 13 with the above expressions for ϵ_n , ϵ'_n , and plotted on the same figure. One can see that the lower the σ , the better the approximation of $h(Y)$ by its upper and lower bounds. On the contrary, when σ increases, the difference between the entropy and its bounds tend to increase, which seems natural. These differences however can be seen to tend towards a constant for $\sigma \rightarrow \infty$. This can be explained as follows. When σ is large, p is no longer multimodal and tends to the Gaussian density of variance σ^2 . Thus $h(Y)$ grows with σ as $\log \sigma$. On the other hand, the upper bound $\mathcal{H}[p]$ on $h(Y)$ also grows as $\log \sigma$. The same is true for the lower bound on $h(Y)$ which equals $\mathcal{H}[p] - \sum_{n=1}^3 \pi_n [\epsilon'_n + \epsilon_n (\log \pi_n^{-1} + 1)]$: the last term tends to $H[\pi] + \frac{3}{2}$ as $\sigma \rightarrow \infty$ since for fixed d_n , $\epsilon_n \rightarrow 1$ and $\epsilon'_n \rightarrow 1/2$ as $\sigma \rightarrow \infty$.

Remark 18 (Entropy bounds and decision theory) The entropy estimator given in Eq. (3.7) has actually close connections with decision problems, and a tighter upper bound for $h[p]$ can be found in this framework, even if the intuitive perception of the approximator in terms of density estimation is lost. The bounds comparison requires to use the base 2 logarithm, i.e. to express entropies in bits. Assume we have an N -class classification problem consisting in finding the class label of an observation y_n knowing the densities and the priors of the classes. In such kinds of classification problems, one is often interested in quantifying Bayes probability of error $P(e)$ which is always positive. In our framework, each of the pdf modes K_n represents the density of a given class c_n , that is $\Pr(Y \leq y | C = c_n) = \int_{-\infty}^y K_n(x)dx$, π_n is the a priori probability of c_n : $\Pr(C = c_n) = \pi_n$ and $p(y)$ is the density associated to the model describing Y , from which the sample y_n is drawn. Denoting the equivocation (which is nothing but the expectation of the conditional entropy) of Y given C by $h(Y|C) = E_C[h(Y|C = c_i)]$ and defining $H(C) \doteq -\sum_{n=1}^N \Pr(C = c_n) \log \Pr(C = c_n)$, it can

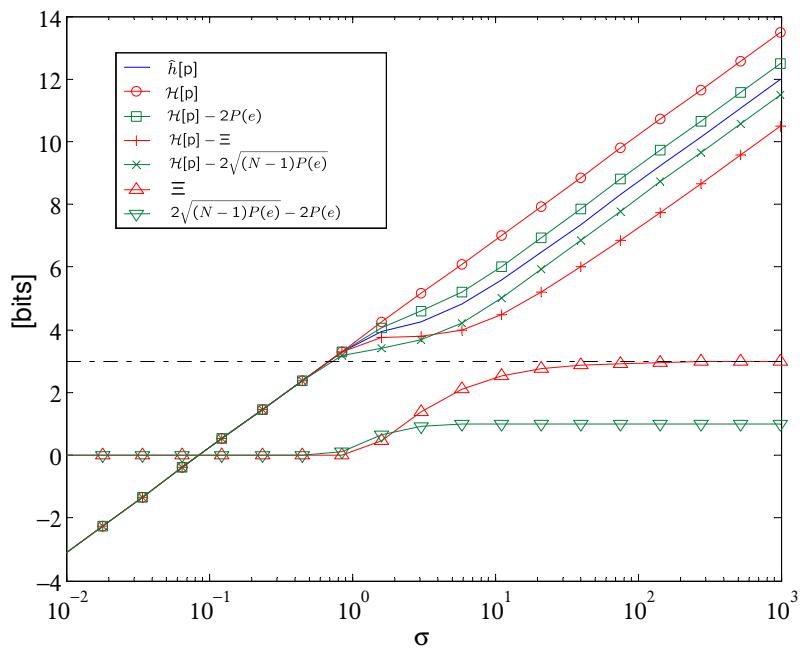


Figure 3.16. Example 15. Evolution of $\bar{h}[p]$ and its bounds versus σ , where if $Y \sim p$, $Y = U + \sigma Z$, U is a discrete random variable taking values in $\{0, 5, 10\}$ with probabilities $\pi = [1/4, 1/2, 1/4]$ and Z is a standard Gaussian variable independent from U . The lower bound $H[p] - \Xi$ converges to the upper bound $H[p]$ and the lower bound $H[p] - 2\sqrt{(N-1)P(e)}$ converges to the upper bound $H[p] - 2P(e)$ as $\sigma \rightarrow 0$ and Ξ tends to $3/2 + H(\pi)$, respectively, as $\sigma \rightarrow \infty$ (note that the horizontal axis scale is logarithmic).

be shown [M. Hellman, 1970, Lin, 1991] that

$$\begin{aligned} P(e) &\leq \frac{1}{2} H(C|Y) \\ &= \frac{1}{2} (h(Y|C) + H(C) - h(Y)) \\ &= \frac{1}{2} \left(\sum_{n=1}^N \pi_n h[K_n] + H[\boldsymbol{\pi}] - h[p] \right) \end{aligned} \quad (3.38)$$

which shows that (half) the difference between $H[p]$ and $h[p]$ is precisely an upper bound on Bayes' probability of error $P(e) \doteq E_Y[1 - \max_i p(C = c_i|y)]$. The error vanishes when the approximator matches to the true value, that is when the approximations (a)-(d) in Eq. (3.7) hold true, in other words, when the modes have no overlap (the classes are separable, i.e. disjoint).

Clearly, $H[p] - 2P(e)$ is a tighter upper bound on $h[p]$ than $H[p]$ as $P(e) \geq 0$. The Generalized Jensen-Shannon divergence also provides a lower bound on the Bayes probability of error. Indeed, it can be proved that $H[p] - 2\sqrt{(N-1)P(e)}$ lower-bounds $h[p]$ [Lin, 1991]. However, the bound in (3.32) is tighter when σ is small enough (see Fig. 3.16.). Note that $P(e)$ is easy to compute : $P(e) = 1 - \int_y p(y) \max_i p(C = c_i|y) dy$ which reduces, using Bayes' rule to $1 - \int_y \max_n \pi_n K_n(y) dy$. In the Gaussian kernel case, this integral can easily be computed without any approximation by successive integrations, whose theoretical expression can be found via the complementary error function Erfc. The last integration bounds are given by the intersections between the Gaussian kernels. However, even if both couples of bounds have a similar computational complexity when applied to a specific example (fixed parameters) and can be theoretically computed, the bounds given in Lemma 13 are easier to deal with in more general theoretical developments, have closer relationship to the multimodality of $p(y)$ and suffice for our purposes. Therefore, in the following theoretical developments, the lower bound given by the right-hand side of Eq. (3.32) and the upper bound given by Eq. (3.31) shall be used.

3.2.3.2 Mixing local minima in multimodal BSS

Based on the results derived in Section 3.2.3.1, it will be shown that mixing local minima of the entropy exist in the context of the blind separation of multimodal sources with Gaussian modes if the mode standard deviations σ_n are small enough.

We are interested in the (mixing) local minima of $h(\mathbf{wS})$ on the K -dimensional unit sphere $\mathcal{S}(K)$ (see Eq. (1.95)). We shall assume that the sources have a pdf of the form (3.2), with K_n being Gaussian with identical standard deviation σ (but with distinct means). Thus, as in Example 15, we may represent S_k as $U_k + \sigma Z_k$ where U_k is a discrete random variable and Z_k is a standard Gaussian variable independent from U_k . Further, $(U_1, Z_1), \dots, (U_K, Z_K)$ are assumed to be independent so that the sources are independent as required. From this

representation, $\mathbf{w}\mathbf{S} = \mathbf{w}\mathbf{U} + \sigma\mathbf{Z}$ where \mathbf{U} is the column vector with components U_k and \mathbf{Z} is again a standard Gaussian variable (since any linear combination of independent Gaussian variables is a Gaussian variable and $\mathbf{Z} = \sum_{k=1}^K w_k Z_k$ has zero mean and unit variance if $\mathbf{w} \in \mathcal{S}(K)$). Since $\mathbf{w}\mathbf{U}$ is clearly a discrete random variable, $\mathbf{w}\mathbf{S}$ also has a multimodal density of the form (3.2) with K_n again the Gaussian density with standard deviation σ . Note that the number of modes is the number of distinct values $\mathbf{w}\mathbf{U}$ can have and the mode centers (the means of the K_n) are these values; they depend on \mathbf{w} . Anyway, as long as σ is small enough with respect to the distances d_n defined in (3.36) the approximation (3.7) of the entropy is justified. Thus, we are led to the approximation

$$h(\mathbf{w}\mathbf{S}) \approx H(\mathbf{w}\mathbf{U}) + \log \sigma + h[\phi], \quad (3.39)$$

where $H(\mathbf{w}\mathbf{U})$ denotes the entropy of the discrete random variable $\mathbf{w}\mathbf{U}$ (remind that the entropy of a discrete random variable \mathbf{U} with probability vector π is noted either $H(\mathbf{U})$ or $H[\pi]$).

The above approximation suggests that there is a relationship between the local minimum points of $h(\mathbf{w}\mathbf{S})$ and those of $H(\mathbf{w}\mathbf{U})$. Therefore, we shall first focus on the local minimum points of the (discrete) entropy of $\mathbf{w}\mathbf{U}$ before analyzing those of $h(\mathbf{w}\mathbf{S})$.

3.2.3.3 Local minimum points of $H(\mathbf{w}\mathbf{U})$

From the definition of discrete entropy in Eq. (1.72), it is seen that $H(\mathbf{w}\mathbf{U})$ does not depend on the values that $\mathbf{w}\mathbf{U}$ can take but only on the associated probabilities; these probabilities remain constant as \mathbf{w} changes unless the number of distinct values that $\mathbf{w}\mathbf{U}$ can take varies. This number would decrease when an equality $\mathbf{w}\mathbf{u} = \mathbf{w}\mathbf{u}'$ is attained for some distinct column vectors \mathbf{u} and \mathbf{u}' in the set \mathcal{U} of possible values of \mathbf{U} . A deeper analysis yields the following result, which is helpful to find the local minimum point of $H(\mathbf{w}\mathbf{U})$.

Lemma 14 *Let \mathbf{U} be a discrete random vector in \mathbb{R}^K and let \mathcal{U} be the set of distinct values it can take. Assume that there exists $r \geq 1$ disjoint subsets $\mathcal{U}_1, \dots, \mathcal{U}_r$ of \mathcal{U} each containing at least 2 elements, such that the linear subspace V spanned by the vectors $\mathbf{u} - \mathbf{u}_1, \mathbf{u} \in \mathcal{U}_1 \setminus \{\mathbf{u}_1\}, \dots, \mathbf{u} - \mathbf{u}_r, \mathbf{u} \in \mathcal{U}_r \setminus \{\mathbf{u}_r\}$, $\mathbf{u}_1, \dots, \mathbf{u}_r$ being arbitrary elements of $\mathcal{U}_1, \dots, \mathcal{U}_r$, is of dimension $K - 1$. (Note that V does not depend on the choice of $\mathbf{u}_1, \dots, \mathbf{u}_r$, since $\mathbf{u} - \mathbf{u}'_j = (\mathbf{u} - \mathbf{u}_j) - (\mathbf{u}'_j - \mathbf{u}_j)$ for any other $\mathbf{u}'_j \in \mathcal{U}_j$.) Then for $\mathbf{w}^* \in \mathcal{S}(K)$ and orthogonal to V , there exists a neighborhood \mathcal{W} of \mathbf{w}^* in $\mathcal{S}(K)$ and $\alpha > 0$ such that $H(\mathbf{w}\mathbf{U}) \geq H(\mathbf{w}^*\mathbf{U}) + \alpha$ for all $\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^*\}$. In the case $K = 2$, one has a stronger result that $H(\mathbf{w}\mathbf{U}) = H(\mathbf{U}) > H(\mathbf{w}^*\mathbf{U})$ for all $\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^*\}$.*

The proof is given in the Appendix at the end of the Chapter, in Section 3.8.5, p. 173.

Example 16 *An illustration of Lemma 14 in the $K = 2$ case (again for clarity) is provided in Fig. 3.17. We note $\mathbf{U} = [U_1, U_2]^T$ where the discrete variables U_1*

and U_2 take the values $-\sqrt{1.03} + 2.5, \sqrt{1.03} + 2.5$ with probabilities .5,.5 and the values $-1.2, -0.4, 2$ with probabilities $1/2, 3/8, 1/8$, respectively. For this random vector U :

$$\begin{aligned} \mathcal{U} = & \left\{ \begin{bmatrix} -\sqrt{1.03} + 2.5 \\ -1.2 \end{bmatrix}, \begin{bmatrix} -\sqrt{1.03} + 2.5 \\ -0.4 \end{bmatrix}, \begin{bmatrix} -\sqrt{1.03} + 2.5 \\ 2 \end{bmatrix} \right\} \\ & \cup \left\{ \begin{bmatrix} \sqrt{1.03} + 2.5 \\ -1.2 \end{bmatrix}, \begin{bmatrix} \sqrt{1.03} + 2.5 \\ -0.4 \end{bmatrix}, \begin{bmatrix} \sqrt{1.03} + 2.5 \\ 2 \end{bmatrix} \right\} \end{aligned}$$

The parameters of U are chosen to have the same variance, as we need that the $S_k = U_k + \sigma Z_k$, $k = 1, 2$, have the same variance. But their mean can be arbitrary since $H(wS)$ does not depend on them. In this $K = 2$ example, each line that links two distinct points $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$ spans a one dimensional linear subspace, which constitutes a possible subspace V , as stated in Lemma 14. There are thus many possibilities for V , and for each of them, a corresponding vector w^* can be found.

Two simple possibilities for V are the subspaces with direction given by $[0, 1]^T$ and $[1, 0]^T$. In the first case, the subsets \mathcal{U}_i are built by grouping the points of \mathcal{U} laying on a same vertical dashed line. There are two such subsets ($r = 2$) consisting of $\mathbf{u} \in \mathcal{U}$ with first component equal to $-\sqrt{1.03} + 2.5$ and $\sqrt{1.03} + 2.5$, respectively. In the second case, the subsets \mathcal{U}_i are built by grouping the points of \mathcal{U} laying on a same horizontal dashed line. There are three such subsets ($r = 3$) consisting of $\mathbf{u} \in \mathcal{U}$ with second component equal to $-1.2, -0.4$ and 2 , respectively.

There also exist other subspaces V , corresponding to “diagonal lines” (i.e. to solid lines in Fig.3.17.). This last kind of one-dimensional linear subspace V correspond to directions given by two-dimensional vectors w^* with two non-zero elements.

On the plot, the points on the half unitary circle correspond to the vectors w^* of the lemma; each w^* is orthogonal to a line joining a pair of distinct points in \mathcal{U} , \mathcal{U} being the set of all possible values of $[U_1, U_2]^T$. The points of \mathcal{U} are displayed in the plot together with their probabilities. The entropies $H(wU)$ are also given in the plot; one can see that they are lower for $w = w^*$ than for other points w .

The above lemma only provides a way to find a local minimum point of the function $H(wU)$, but does not prove the existence of such a point, since the existence of V was only assumed in the lemma. Nevertheless, in the case where the components of U are independent and can take at least 2 distinct values, subset \mathcal{U}_i ensuring the existence of V can be built as follows. Let j be any index in $\{1, \dots, K\}$ and let $\lambda_{j,1}, \dots, \lambda_{j,r_j}$ be the possible values of the j -th component of U . One can take $\mathcal{U}_i, 1 \leq i \leq r_j$ to be the set of vectors $\mathbf{u} \in \mathcal{U}$ such that their j -th components $\mathbf{u}(j)$ equal $\lambda_{j,i}$. Considering Example 16, we have $\lambda_{1,1} = -\sqrt{1.03} + 2.5$ and $\lambda_{1,2} = \sqrt{1.03} + 2.5$ ($r_1 = 2$) and $\lambda_{2,1} = -1.2, \lambda_{2,2} = -0.4, \lambda_{2,3} = 2$ ($r_2 = 3$).

Then it is clear that the corresponding subspace V consists of all vectors orthogonal to e_j , i.e. the j -th row of the identity matrix (hence V is of dimension $K - 1$) and that the associated vector w^* is simply this row or its opposite.

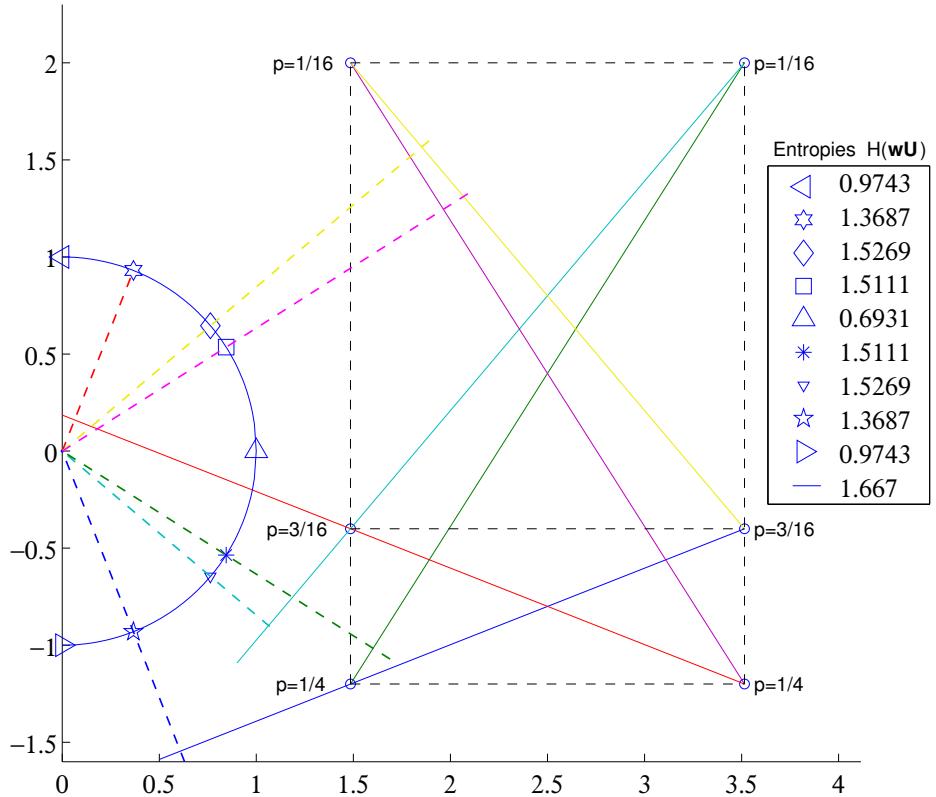


Figure 3.17. Example 16. Illustration of Lemma 14. The discrete random variables U_1 and U_2 take values in $\{-\sqrt{1.03}+2.5, \sqrt{1.03}+2.5\}$ and $\{-1.2, -.4, 2\}$ with probabilities $[.5.5]$ and $[1/2, 3/8, 1/8]$, respectively. The entropies at the points located by the corresponding markers shown on the half-circle are given in the legend. ©2006, IEEE. Reprinted, with permission, from Vrins, Pham & Verleysen: *Mixing and non-mixing local minima of the entropy contrast for blind source separation. To appear in IEEE Transactions on Information Theory (March 2007)*.

Observe that for $(\mathbf{u}, \mathbf{u}') \in \mathcal{U}_i$, $\mathbf{e}_j \mathbf{u} = \mathbf{u}(j) = \mathbf{u}'(j) = \lambda_{j,i}$, leading to $\mathbf{e}_j(\mathbf{u} - \mathbf{u}') = 0$. By Lemma 14, this point \mathbf{w}^* would be a local minimum point of $H(\mathbf{w}\mathbf{U})$. But, as explained above, it is a non mixing point while we are interested in the mixing points, i.e. not proportional to a row of the identity matrix. However, the above construction can be extended by looking for a set of K vectors $\mathbf{u}_1, \dots, \mathbf{u}_K$ in \mathcal{U} , such that the vectors $\mathbf{u}_i - \mathbf{u}_j, 1 \leq i < j \leq K$ span any linear subspace of dimension $K - 1$ of \mathbb{R}^K . If such a set can be found, then V is simply this linear subspace by taking $\mathcal{U}_1 = \{\mathbf{u}_1, \dots, \mathbf{u}_K\}$ and $r = 1$. In addition, if $\mathbf{u}_1, \dots, \mathbf{u}_K$ do not all have the same j -th component, for some j , then the corresponding \mathbf{w}^* is a mixing local minimum point. In view of the fact that there are at least 2^K points in \mathcal{U} to choose from for the \mathbf{u}_i and that the last construction procedure meant not find all local minimum points of $H(\mathbf{w}\mathbf{U})$, chance is that there exists both non-mixing and mixing local minimum points of $H(\mathbf{w}\mathbf{U})$. In the $K = 2$ case this is really the case: it suffices to take two distinct points \mathbf{u}_1 and \mathbf{u}_2 in \mathcal{U} , then by the above lemma, the vector \mathbf{w}^* orthogonal to $\mathbf{u}_1 - \mathbf{u}_2$ is a local minimum point of $H(\mathbf{w}\mathbf{U})$. If one chooses \mathbf{u}_1 and \mathbf{u}_2 such that both components of $\mathbf{u}_1 - \mathbf{u}_2$ are non zero, the associated orthogonal vector \mathbf{w}^* is not proportional to any row of the identity matrix; it is a mixing local minimum point of $H(\mathbf{w}\mathbf{U})$. In Example 16, we can take $\mathbf{u}_1 = [-\sqrt{1.03} + 2.5, -1.2]^T$ and $\mathbf{u}_2 = [\sqrt{1.03} + 2.5, 2]^T$. Note that in the particular $K = 2$ case, the aforementioned method identifies all local minimum points of $H(\mathbf{w}\mathbf{U})$. Indeed, for any $\mathbf{w} \in \mathcal{S}(K)$, either there exists a pair of distinct vectors $\mathbf{u}_1, \mathbf{u}_2$ in \mathcal{U} such that $\mathbf{w}(\mathbf{u}_1 - \mathbf{u}_2) = 0$ or there exists no such pair. In the first case \mathbf{w} is a local minimum point and in the second case one has $H(\mathbf{w}\mathbf{U}) = H(\mathbf{U})$. Since there is only a finite number of the $\mathbf{u}_1 - \mathbf{u}_2$ differences, for distinct $\mathbf{u}_1, \mathbf{u}_2$ in \mathcal{U} , there can be only a finite number of local minimum points of $H(\mathbf{w}\mathbf{U})$; for all other points $H(\mathbf{w}\mathbf{U})$ take the maximum value $H(\mathbf{U})$.

Remark 19 (Relationship between informal and formal approaches)

Let us focus on $K = 2$ for simplicity. The informal approach states that local entropy minima appear when the elements of the transfer vectors are such that two intermodal distances of the scaled source pdfs become equal. In other words, when $w_1 \Delta_{ij}(\mathbf{S}_1) = w_2 \Delta_{pq}(\mathbf{S}_2)$. The fact is that the values that the k -th component of the vectors \mathbf{u} can take correspond to the modes of the density of \mathbf{S}_k : therefore, two intermodal distances of the k -th source are given by $\mathbf{u}(k) - \mathbf{u}'(k)$ for some \mathbf{u}, \mathbf{u}' with different k -th component. Two intermodal distances of the scaled source pdf will be equal if $w_1(|\mathbf{u}(1) - \mathbf{u}'(1)|) = w_2(|\mathbf{u}(2) - \mathbf{u}'(2)|)$; this will be the case if $w_1(\mathbf{u}(1) - \mathbf{u}'(1)) = -w_2(\mathbf{u}(2) - \mathbf{u}'(2))$, i.e. when $\mathbf{w} = \mathbf{w}^$ with $\mathbf{w}^*(\mathbf{u} - \mathbf{u}') = 0$. Noting $\mathbf{w}^* = \mathbf{w}_{\theta^*}$, this indicates that a local minimum exists when $\theta = \arctan \left(\frac{|\mathbf{u}(2) - \mathbf{u}'(2)|}{|\mathbf{u}(2) - \mathbf{u}'(2)|} \right)$, i.e. when $\theta = \pm \theta^*$ or $\theta = \pi \pm \theta^*$.*

3.2.3.4 Local minimum points of $h(\mathbf{w}\mathbf{S})$

Equation (3.39) suggested that the local minimum points of $h(\mathbf{w}\mathbf{S})$ are closely related to those of $H(\mathbf{w}\mathbf{U})$. Therefore, the entropy local minima of the discrete

random variable \mathbf{wU} have been analyzed in the above subsection. Based on the approximation (3.39), this result extends to $h(\mathbf{wS})$ for sufficiently small σ . But the problem is that Eq. (3.39) remains vague : we don't know how close $h(\mathbf{wS})$ is from its upper bound $\mathcal{H}(\mathbf{wS})$ in the neighborhood of the local minimum points¹.

In a complementary way, the following Lemma 15 finally relates formally the local minimum points of $H(\mathbf{wU})$ to those of $h(\mathbf{wS})$. A last example is proposed to illustrate the theoretical results.

Lemma 15 *Define S_i , $i = 1, \dots, K$, as $S_i = U_i + \sigma Z_i$ described at the beginning of subsection 3.2.3.2 and let \mathbf{w}^* be a vector satisfying the assumption of Lemma 14 (\mathbf{U} being the vector with component U_i). Then for σ sufficiently small $h(\mathbf{wS})$ admits a local minimum point converging to \mathbf{w}^* as $\sigma \rightarrow 0$.*

The proof of this lemma is relegated to the Appendix of the Chapter, in Section 3.8.6 (p. 174).

Example 17 *Thanks to the entropy approximator, we shall illustrate the existence of the local minima of $h(\mathbf{wS})$ in the following $K = 2$ example, so that vectors $\mathbf{w} \in \mathcal{S}(K)$ can be written, as usual, as $\mathbf{w}_\theta = [\sin \theta, \cos \theta]$. We take $S_1 = U_{\pi/2} + \sigma Z_1$ and $S_2 = U_0 + \sigma Z_2$, where $U_0, U_{\pi/2}$ are independent discrete random variables taking the values $-2\sqrt{3}/3, \sqrt{3}/2$ with probabilities $1/3, 2/3$ and $-\sqrt{2}, \sqrt{2}/2$ with probabilities $3/7, 4/7$, respectively, and Z_1, Z_2 are standard Gaussian variables. The source pdf are then different and asymmetric, contrarily to the assumptions drawn in Section 3.2.2. The parameter σ is set to 0.1. Thus $Y_\theta = \mathbf{w}_\theta S$ can be represented as $U_\theta + \sigma Z$ where $U_\theta = \sin \theta U_{\pi/2} + \cos \theta U_0$ and Z is a standard Gaussian variable independent from U_θ . Figure 3.18. plots the pdf of Y_θ for various angles θ . It can be seen that the modality (i.e. the number of modes) changes with θ . Fig. 3.19. shows the entropy $\bar{h}(Y_\theta)$ together with its upper and lower bounds, for $\theta \in [0, \pi]$. In addition to non-mixing local minima at $\theta \in \{p\pi/2 | p \in \mathbb{Z}\}$, mixing local minima exist when $\mathbf{w}_\theta(\mathbf{u}_1 - \mathbf{u}_2) = 0$, where $\mathbf{u}_1 = [-2\sqrt{3}/3, \sqrt{2}/2]^T$, $\mathbf{u}_2 = [\sqrt{3}/2, -\sqrt{2}]^T$, i.e. when $|\tan(\theta)| = .9526$, or $\theta \in \{(0.2423 + p)\pi, (0.7577 + p)\pi | p \in \mathbb{Z}\}$. One can observe that the upper bound is a constant function except for a finite number of angles for which we observe negative peaks (see Lemma 14). For these angles the pdf is strongly multimodal, and the upper and lower bounds are very close, though not clearly visible on the figure. This results from a discontinuity of the lower and upper entropy bounds at these angles, due to the superimposition of several modes.*

¹We have used, as usual, the shorthand notation $\mathcal{H}(\mathbf{wS}) = \mathcal{H}[p]$ where p is the pdf of the random variable \mathbf{wS}

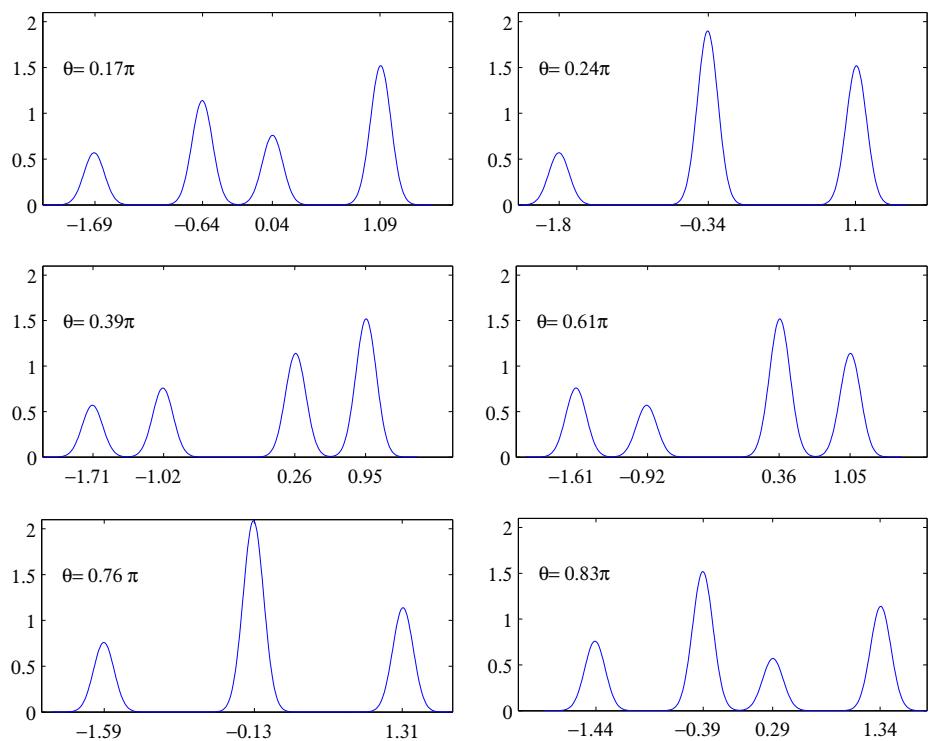


Figure 3.18. Example 17. Probability density function of $w_\theta S$ for various angles θ . ©2006, IEEE. Reprinted, with permission, from Vrins, Pham & Verleysen: *Mixing and non-mixing local minima of the entropy contrast for blind source separation*. To appear in *IEEE Transactions on Information Theory* (March 2007).

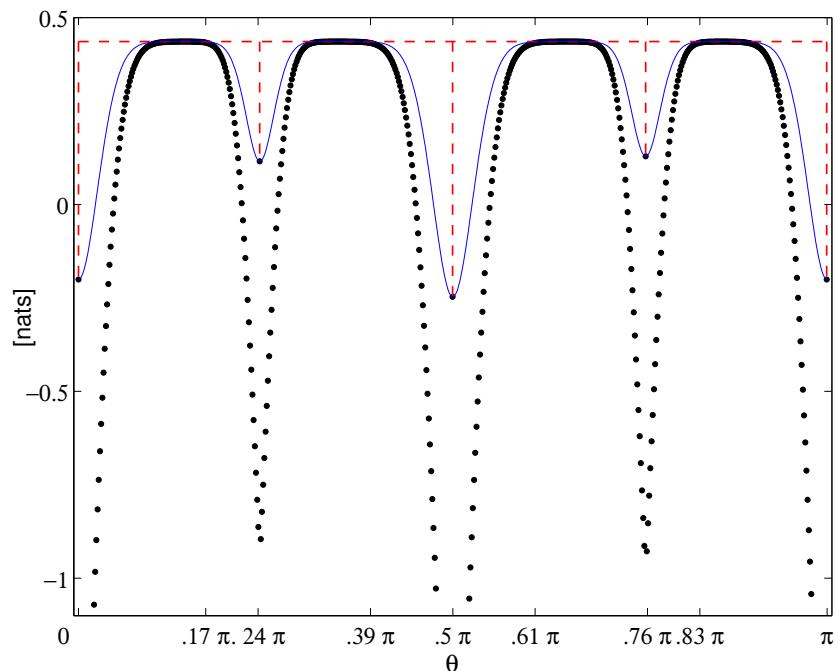


Figure 3.19. Example 6. Upper bound (dashed), lower bound (dots) and entropy estimation $\bar{h}(Y_\theta)$ (solid). The upper and lower bounds of the entropy converge to each other when the density becomes strongly multimodal because there are downwards and upwards jumps of $H[p]$ and $H[p] - \Xi$, respectively (see also the corresponding plots in Fig. 3.18.). ©2006, IEEE. Reprinted, with permission, from Vrins, Pham & Verleysen: *Mixing and non-mixing local minima of the entropy contrast for blind source separation*. To appear in IEEE Transactions on Information Theory (March 2007).

3.2.3.5 Complementary observations

This section provides two observations that can be drawn regarding the impact of the *mode variance* σ^2 on the existence of local minima and the symmetry of the entropy with respect to θ .

- Impact of “mode variance” σ^2

In the example of Fig. 3.20, the discrete variables U_1 and U_2 in the expression of S_1 and S_2 are taken as in Example 16. One can observe that the mixing minima of the entropy tends to disappear when the mode variance increases. This is a direct consequence of the fact that the mode overlaps increase. When σ increases, the source densities become more and more Gaussian and the $h(Y_\theta)$ vs θ curve tends to be more and more flat, approaching the constant function $\log \sqrt{2\pi e} + \log \sigma$. The upper and lower bounds have only been plotted for the $\sigma = .05$, for visibility purposes. Again, at angles corresponding to the upper bound negative peaks, the error bound is very tight, as explained in Example 17.

- Note on the symmetry of $h(Y_\theta)$

In the above graphs plotting the (Riemannian estimated) entropy (and its bounds) versus θ , some symmetry of h can be observed. First, observe that $h(Y_\theta) = h(Y_{\theta+\pi})$ whatever the source pdfs; this is a direct consequence of the fact the the entropy is not sign sensitive. Second, if one of the source densities is symmetric, i.e. if there exists $\mu \in \mathbb{R}$ such that $p_{S_j}(\mu - s) = p_{S_j}(\mu + s)$ for all $s \in \mathbb{R}$, then $h(Y_\theta) = h(Y_{-\theta})$. Third, if the two sources share the same symmetric pdf, then $h(Y_\theta) = h(Y_{\pi/2-\theta})$. Finally, if the two sources can be expressed as in Lemma 15, then the vectors \mathbf{w}_θ^* for which $H(\mathbf{w}_\theta^* U) < H(U)$ (as obtained in Lemma 14) are symmetric in the sense that their angles are pairwise opposite. This means that for σ small enough, if a local minimum of $h(\mathbf{w}_\theta S)$ appears at θ^* , then another local minimum point will exist near $-\theta^*$ (and thus near $p\pi - \theta^*$, $\forall p \in \mathbb{Z}$). The above symmetry property can be seen from Figure 3.17. and can be proved formally as follows. From Lemma 14, \mathbf{w}^* must be orthogonal to $\mathbf{u}_1 - \mathbf{u}_2$ for some pair of distinct vectors in the set of all possible values of U . Define \mathbf{u}_i^\dagger ($i = 1, 2$) to be the vector with first coordinate the same as that of \mathbf{u}_{3-i} and second coordinate the same as that of \mathbf{u}_i . Then it can be seen that the vector orthogonal to $\mathbf{u}_1^\dagger - \mathbf{u}_2^\dagger$ has an angle opposite to the angle of \mathbf{w}^* , yielding the desired result. We recover the results of the informal analysis made in Rem. 19.

3.2.4 Cumulant-based versus Information-theoretic approaches

Consider two pairs of independent, whitened, and multimodal sources with non-zero kurtosis. For zero-mean and unit variance signals, the kurtosis $\kappa(Y)$ is nothing else than the fourth order auto-cumulant (see Eq. (1.28)). In the first pair of sources, both are bimodal ($N(S_1) = N(S_2) = 2$; their analytical form

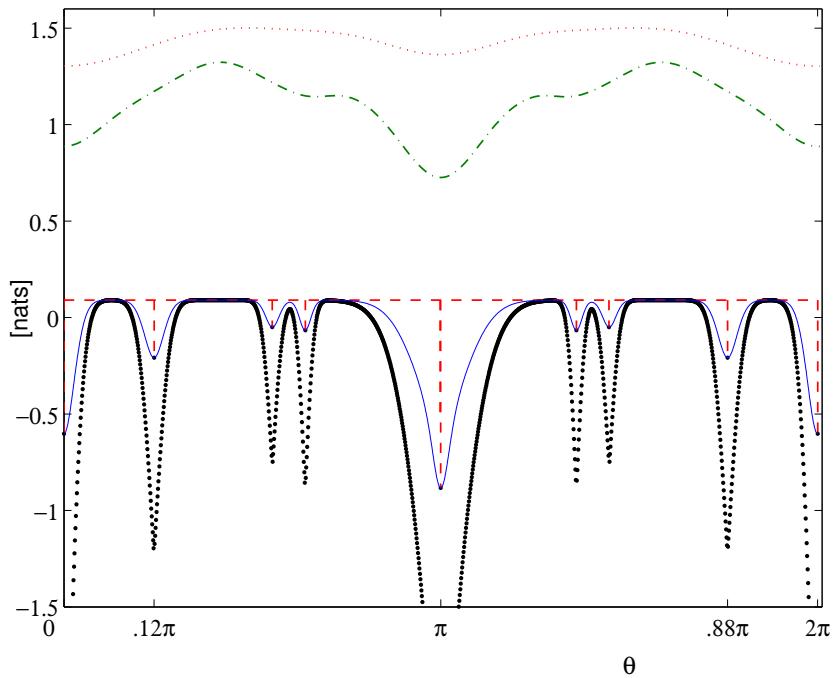


Figure 3.20. Entropy $\bar{h}(Y_\theta)$ versus θ for $S_1 = U_1 + \sigma Z_1$, $S_2 = U_2 + \sigma Z_2$, where U_1 and U_2 are taken from Example 16 (and Fig. 3.17.) and the four random variables are all independent. The parameter σ is set to .05 (solid), .25 (dash-dotted) and .5 (dotted). The upper and lower bounds have been added for the $\sigma = .05$ case only, for visibility purposes. The upper and lower bounds of the entropy converge to each other when the density becomes strongly multimodal. ©2006, IEEE. Reprinted, with permission, from Vrins, Pham & Verleysen: *Mixing and non-mixing local minima of the entropy contrast for blind source separation. To appear in IEEE Transactions on Information Theory*.

does not matter up to the above specificities), while in the second pair, the first is bimodal and the second trimodal ($N(S_1) = 2, N(S_2) = 3$). We shall analyze the evolution of $\mathbf{Y} = \mathbf{R}_\theta \mathbf{S}$ where θ is the transfer angle and \mathbf{R}_θ a 2D matrix rotation. Figure 3.21.(a) and Figure 3.21.(b) show $\hat{h}(Y_1)$ and the output mutual information $\hat{h}(Y_1) + \hat{h}(Y_2)$ for both pairs of sources. The same is done for $\kappa^2(Y_1)$ and $\kappa^2(Y_1) + \kappa^2(Y_2)$ (Fig. 3.22.(a) and Fig. 3.22.(b)). It can be seen that while the entropic criterion suffers from spurious minimum points, the above kurtotic criteria do not. Theoretical proofs regarding the spurious minimum points of entropic criteria were given in the above sections of this chapter. Under the whiteness constraint, the discriminacy property of $\kappa^2(Y_1)$ has been proved in [Delfosse and Loubaton, 1995] while in the specific 2D case (that is when angular parametrization of the transfer matrix is possible), [Murillo-Fuentes and Gonzalez-Serrano, 2004] proved the same discriminacy property of $\kappa^2(Y_1/\sqrt{\text{Var}[Y_1]}) + \kappa^2(Y_2/\sqrt{\text{Var}[Y_2]})$ under the $\text{Cov}[\mathbf{Y}] = \mathbf{I}_2$ constraint.

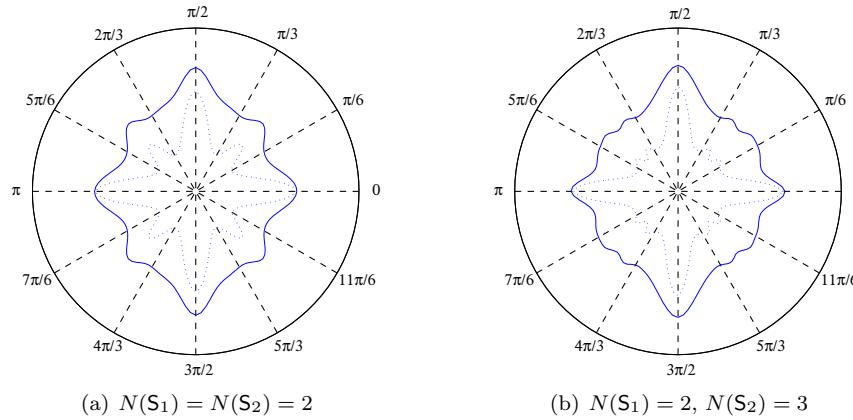


Figure 3.21. Plot of $-\hat{h}(Y_1)$ (solid) and of $-\hat{h}(Y_1) - \hat{h}(Y_2)$ (dotted) vs θ for a pair of bimodal sources (a) and for a mixture of bimodal and trimodal sources (b). Mixing maxima are seen. ©2005, IEEE. Reprinted, with permission, from Vrins & Verleysen: *Information theoretic vs cumulant-based contrasts for multimodal source separation*. IEEE Signal Processing Letters 12(3), pp. 190-193, March 2005.

In Section 3.2.1.1, it is explained that $N(Y_1)$ may vary between $\min(N(S_1), N(S_2))$ and $N(S_1) \cdot N(S_2)$ for θ varying between $[k\pi/2, (k+1)\pi/2]$. It is emphasized that $N(Y_1)$, when expressed as a function of θ , may have local minima in $]k\pi/2, (k+1)\pi/2[$; these minima coincide with the (spurious) local minima of $\hat{h}(Y_1)$, i.e. the spurious local maxima of $-\hat{h}(Y_1)$. This can be observed comparing Fig. 3.21.(a) and Fig. 3.23.

The analysis held in this section has been extended to approximations of the entropy other than \hat{h} , such as nearest-neighbor approximators of entropy (spacing estimates of entropy [Learned-Miller and Fisher III, 2003]); the conclusion is identical.

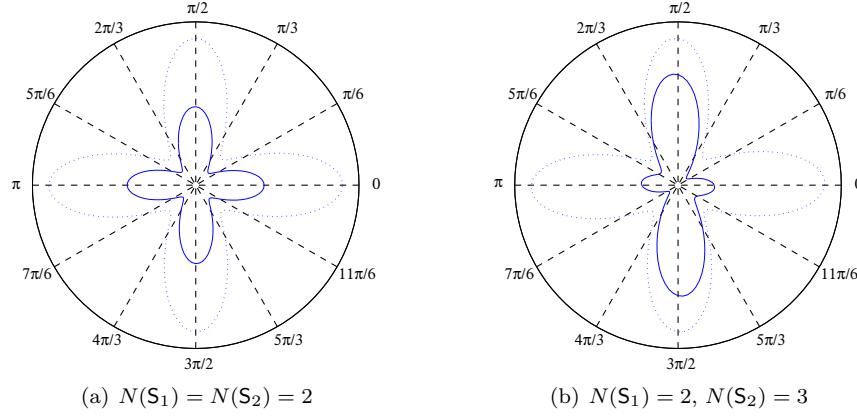


Figure 3.22. Plot of $\kappa^2(Y_1)$ (solid) and of $\kappa^2(Y_1) + \kappa^2(Y_2)$ (dotted) vs θ for a pair of bimodal sources (a) and for a mixture of bimodal and trimodal sources (b); the same as in Fig. 3.21.. No mixing minima exist. ©2005, IEEE. Reprinted, with permission, from *Vrins & Verleysen: Information theoretic vs cumulant-based contrasts for multimodal source separation. IEEE Signal Processing Letters 12(3), pp. 190-193, March 2005.*

Therefore, as both the entropy and the kurtosis depend on the whole density function, an interesting point is to give an intuitive justification for explaining this phenomenon. In other words, why does the kurtosis seem not be sensitive to the modality of the density while entropy does? An element of answer is proposed below.

Information theoretic criteria, as well as cumulant-based ones, map the structure of a density to a real number. Both these criteria measure statistical quantities of densities.

The density of Y_1 is directly related to $p_{S_1}(S_1)$, $p_{S_2}(S_2)$ and θ by Eq. (1.76), and the density resulting of the sum of independent random variables is the convolution of the variable densities.

Comparing Fig. 3.21.(a) and Fig. 3.23., it is clear that $-h(Y)_1$ is a measure of the whole structure of $p_{Y_1}(Y_1)$ (and among others, a function of the number of modes). By contrast, $\kappa^2(Y_1)$ characterizes more specifically the tails of $p_{Y_1}(Y_1)$, discarding its internal structure (in the middle range of the support of Y_1), as visible comparing Fig. 3.23. to Fig. 3.22.(a).

This property of the kurtosis, which can also be used as a non-Gaussianity index [Comon, 1994, Hyvärinen et al., 2001], has been emphasized by J. H. Friedman: (...) projection indexes based on standardized cumulants heavily emphasize the departure from normality in the tails of density. (...) For example, a density with only slightly heavier than normal tails receives a much higher index value than a highly clustered projection (i.e. density) [Friedman, 1987]. This analysis (particularized to the kurtosis) has been translated for the ICA problem in [Hyvärinen and Oja, 1997].

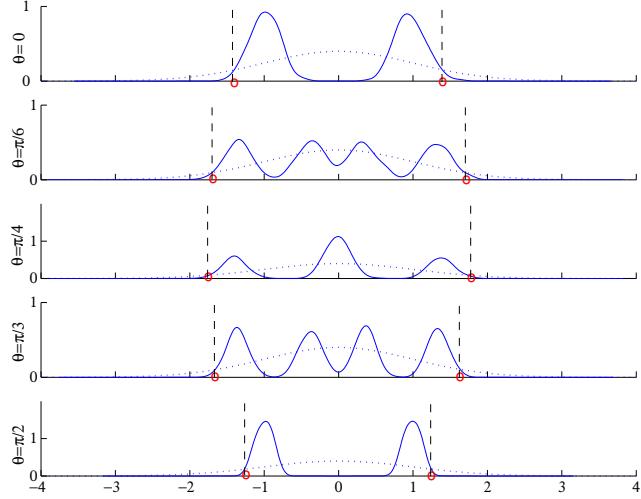


Figure 3.23. Plot of $p_{Y_1}(Y_1)$ for several values of θ (solid) associated to Fig. 3.21.(a) and the standard Gaussian density $\phi(\cdot)$ (dotted). The dashed curves show the leftmost and rightmost intersections between these densities. ©2005, IEEE. Reprinted, with permission, from Vrins & Verleysen: *Information theoretic vs cumulant-based contrasts for multimodal source separation*. IEEE Signal Processing Letters 12(3), pp. 190-193, March 2005.

The previous considerations are illustrated in Fig. 3.22., Fig. 3.23. and Fig. 3.24. In this experiment, $\kappa(Y_1)$ can approximately be seen as a measure of where the tails of $p_{Y_1}(y)$ cross the tails of the Gaussian density $\phi(y)$ of zero mean and unit variance. In other words, if we suppose that $\phi(y) \geq p_{Y_1}(y)$ for $y \geq y_1^{*r}$ and for $y \leq y_1^{*l}$ (with $y_1^{*l} < 0 < y_1^{*r}$), then the lower y_1^{*r} and $|y_1^{*l}|$, the higher $|\kappa(Y_1)|$ (the link between the kurtosis and $|y_1^*|$ can be seen comparing Fig. 3.22. and Fig. 3.24.). The y_1^{*r} and $|y_1^{*l}|$ are indicated by circles in Fig. 3.23. Note that y_1^{*r} or $|y_1^{*l}|$ have similar behavior vs θ . Moreover, as visible in Fig. 3.24., $|y_1^{*l}(\theta)| = y_1^{*r}(\theta + \pi)$ and $y_1^{*r}(\theta) = |y_1^{*l}(\theta + \pi)|$ (since by Eq. (1.59) $Y_i(\theta) = -Y_i(\theta + \pi)$). Hence, the lower the $y_1^* \doteq (y_1^{*r} + |y_1^{*l}|)/2$, the higher the absolute kurtosis $|\kappa(Y_1)|$.

The key point here is to observe that the evolution of $\hat{h}(Y_i)$ is largely influenced by $N(S_1)$ and $N(S_2)$. On the contrary, the evolution of y_1^* versus θ (or more precisely the shape of this function) mainly depends on the transfer coefficients (i.e. on θ); the number $N(Y_1)$ of modes has no influence on the number of extrema of the kurtosis. Even if the source densities may stretch or distort the shape of $|\kappa(Y_1)|$ (expressed as a function of θ), this shape remains similar for both unimodal or multimodal source densities.

Let us define s_i^{*r} , s_i^{*l} and s_i^* similarly to y_1^{*r} , y_1^{*l} and y_1^* , respectively. Starting from $\theta = k\pi$ to $\theta = (2k+1)\pi/2$, y_1^* increases from s_2^* , reaches a (possibly locally) maximum value, and decreases to s_1^* . This is exactly the same scheme as for

unimodal sources separation, and ensures that all locally maximum values of $|\kappa(Y_1)|$ (i.e. the minimum values of y_1^*), that can be detected blindly knowing only the demixing matrix, are attained for $\theta = \{k\pi/2\}$ ($k \in \mathbb{Z}$), corresponding to a non-mixing transfer matrix \mathbf{W}^* . As a consequence, using gradient-based maximization of $\kappa^2(Y_1)$ or $\kappa^2(Y_1) + \kappa^2(Y_2)$ does not lead to spurious solutions. In addition, it is known that algebraic methods can also be used to maximize the last contrast function [Comon, 2001], avoiding spurious solutions too. On the contrary, entropy-based contrast functions are maximized by gradient-based methods; it is shown from experimental and theoretical viewpoints that spurious maxima may appear in this case. The above development appeared in [Vrins and

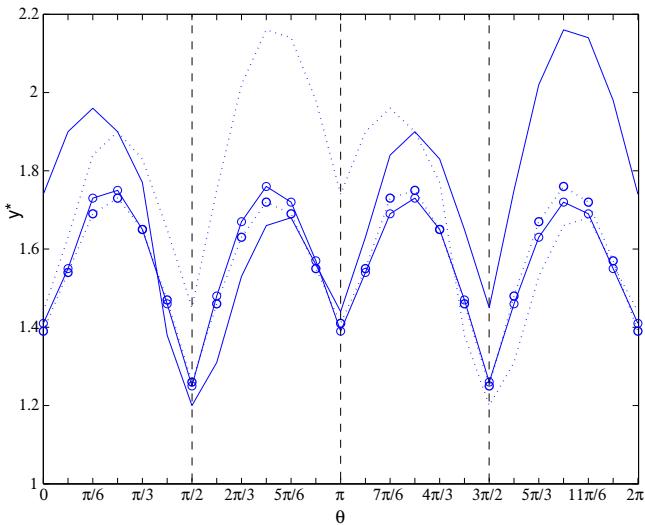


Figure 3.24. Evolution of y_1^{*r} (solid) and $|y_1^{*l}|$ (dotted) versus θ for the examples given in figures 3.22.(a) (markers 'o') and 3.22.(b) (no marker). ©2005, IEEE. Reprinted, with permission, from Vrins & Verleysen: *Information theoretic vs cumulant-based contrasts for multimodal source separation*. IEEE Signal Processing Letters 12(3), pp. 190-193, March 2005.

Verleysen, 2005a].

3.3 DISCRIMINACY OF RÉNYI'S ENTROPY

The discriminacy, as defined in this thesis, is a property of contrast function. However, it has been shown in Chapter 2 (Section 2.4) that Rényi's entropy, generally speaking is not a contrast function, and by corollary, is not a discriminant contrast function. In other words, we have no guarantee that finding the

global maximum point will lead to recover the sources. Consequently, discussing the existence of possible non-mixing maxima is useless and therefore, has not been addressed.

3.4 DISCRIMINACY OF THE MINIMUM RANGE (EXTENDED ZERO-ENTROPY) APPROACH

3.4.1 Preliminaries : $K = 2$ case

Remind the properties of the range functional, summarized in Section 2.3.2. From model given in Eq. (1.61), it comes that

$$-R(Y_\theta) = -|\sin \theta|R(S_1) - |\cos \theta|R(S_2) . \quad (3.40)$$

Let us now turn to the behavior of the above criterion in the first quadrant Q_1 of the unit circle (this result holds for the other quadrants). The second derivative of the above criterion yields

$$-\frac{\partial^2 R(Y_\theta)}{\partial \theta^2} = \cos \theta R(S_1) + \sin \theta R(S_2) > 0 . \quad (3.41)$$

More generally, this inequality ensures that $-R(Y_\theta)$ is convex in θ on each quadrant Q_p and, therefore, the maximum value of $-R(Y_i)$ ($i \in \{1, 2\}$, see model in Eq. (1.60)) is reached for $\theta \in \{k\pi/2\}$:

Corollary 13 *If φ^* is the unmixing angle maximizing locally $-R(Y_\theta)$, then $\theta^* \doteq \phi + \varphi^* = k\pi/2$.*

A typical shape of the contrast function for $K = 2$ and $K = 3$ is given in the next two examples.

Example 18 (Example for $K = 2$) *Figure 3.25. illustrates the contrast function*

$$-1/\|w\|(|w_1|R(S_1) + |w_2|R(S_2))$$

(with $R(S_1) = 2$, $R(S_2) = 3$). The local maximum points only occur at non-mixing vectors. Further, the criterion is seen to be non-sensitive to $\|w\|$, as expected.

Example 19 (Example for $K = 3$) *Suppose $Y = wS$. The $w \in S(3)$ constraint defines a 2-dimensional manifold in \mathbb{R}^3 . Figure 3.26. shows $-R(Y)$ on the $S(3)$ 2-manifold in w_1, w_2 plane ($|w_3| = 1 - w_1^2 - w_2^2$) for two triples of source ranges. The global maxima (darkest areas) correspond to the weight vector $w = e_i$ where $i = \operatorname{argmin}_j R(S_j)$; only the weight corresponding to the source with minimum range is non zero. This weight is equal to one because of the normalization constraint. All the local maxima correspond to $w = e_k$, $1 \leq k \leq K$.*

The above results have been published in [Vrins et al., 2005a].

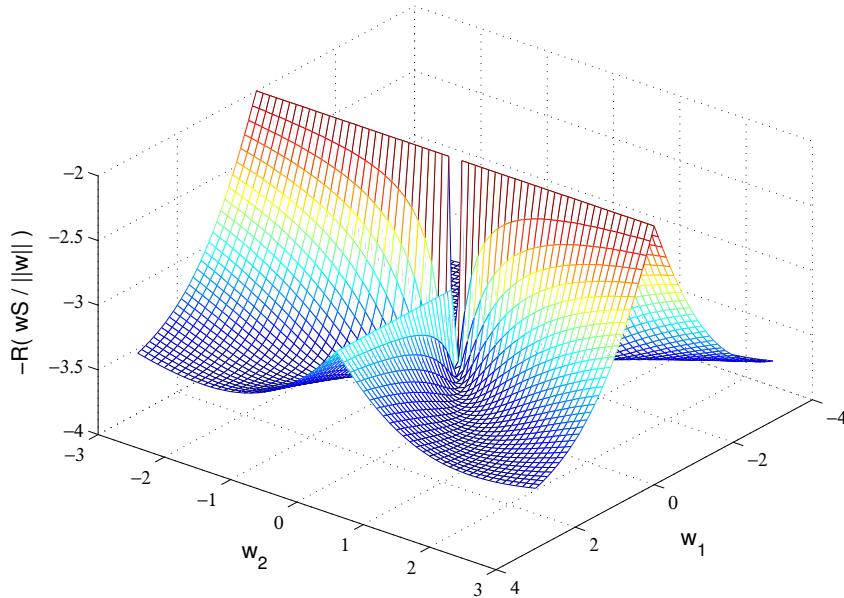


Figure 3.25. Example 18: contrast function $-R(\mathbf{w}\mathcal{S}/\|\mathbf{w}\|)$ on the 2D space w_1, w_2 .

3.4.2 Deflation approach

Proving the discriminacy property of the range-based deflation contrast can be managed by using a small-variation analysis (this approach was presented in [Vrins et al., 2007a]). However, as illustrated in the above introductory section, a “piecewise convex” contrast (i.e. convex between two non-mixing solutions) is necessarily discriminant as it cannot have a maximum point at a mixing (spurious) solution. This more elegant approach will be followed in Section 3.4.2.2 based on the concept of *geodesic convexity*. Finally, a more global approach using the Hessian of the criterion will be proposed; the last approach will be useful when focussing on particular trajectories in the space spanned by the criterion.

3.4.2.1 Using small variation approach and constrained output variance

A simple method for proving the non-existence of spurious maxima in the deflation contrast $\mathcal{C}_R(\mathbf{w})$ under the $\mathbf{w} \in \mathcal{S}(K)$ constraint is to show that at any point $\mathbf{w} \in \mathbb{R}^K$ satisfying $\sharp[I(\mathbf{w})] > 1$, there always exists a small vector $\delta\mathbf{w}$ such that i) $\mathcal{C}_R(\mathbf{w} + \delta\mathbf{w}) > \mathcal{C}_R(\mathbf{w})$, ii) $\|\delta\mathbf{w}\| > 0$ and iii) $\mathbf{w} + \delta\mathbf{w}$ satisfies the constraint $\|\mathbf{w} + \delta\mathbf{w}\| = 1$.

For proving the above result, let us consider the next lemma. We restrict the search space in \mathcal{V}_K^1 as the sign of the entries of \mathbf{w} does not affect the range.

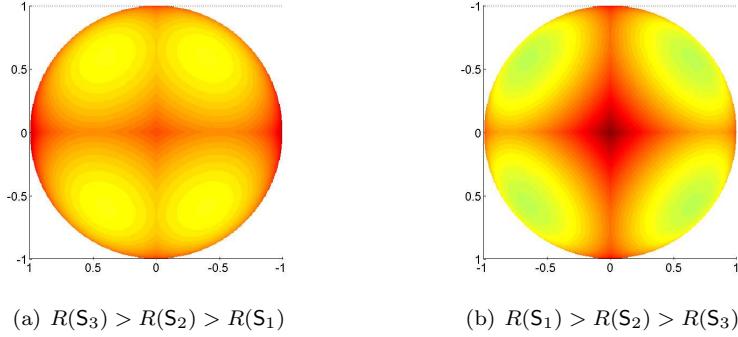


Figure 3.26. Example 19. contrast function $-R(\mathbf{w}\mathbf{S})$ ($\mathbf{w} \in \mathcal{S}(3)$) defined on a 2-manifold, projected on the 2D space w_1, w_2 for a better readability. ©2005, IEEE. Reprinted, with permission, from Vrins, Jutten & Verleysen: *SWM: A class of convex contrasts for source separation*. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. V.161-V.164, March 2005, Philadelphia (USA).

Lemma 16 For all vectors $\mathbf{w} \in \mathcal{V}_K^1$ and two distinct indexes $1 \leq i, j \leq K$ there exists two small scalar numbers ζ, ξ such that if we define $\delta\mathbf{w}_{ij}^\zeta \doteq \zeta\mathbf{e}_i + \xi\mathbf{e}_j$ with

$$\zeta \doteq -\mathbf{w}(j) + \sqrt{\mathbf{w}(j)^2 - (2\mathbf{w}(i)\zeta + \zeta^2)}, \quad (3.42)$$

ensuring that $\mathbf{w} + \delta\mathbf{w}_{ij}^\zeta \in \mathcal{V}_K^1$, then $\|\delta\mathbf{w}_{ij}^\zeta\| < \epsilon$ for all $\epsilon > 0$.

Further, from Eq. (2.50),

$$\Delta\tilde{C}_R(\mathbf{w} + \delta\mathbf{w}_{ij}^\zeta, \mathbf{w}) = R(S_i)\zeta + R(S_j)\xi. \quad (3.43)$$

The proof is straightforward and is given in the Appendix (Section 3.8.7, p. 177).

The above lemma is useful for proving the next key theorem.

Theorem 18 For all $\mathbf{w} \in \mathcal{V}_K^1 \setminus \{\mathbf{e}_1, \dots, \mathbf{e}_K\}$, there exists two distinct indexes $1 \leq i, j \leq K$ such that $0 < \mathbf{w}(i), \mathbf{w}(j) < 1$. For such indexes, consider the small vectors $\delta\mathbf{w}_1, \delta\mathbf{w}_2$ defined as:

$$\begin{aligned} \delta\mathbf{w}_1 &\doteq \delta\mathbf{w}_{ij}^\zeta, \\ \delta\mathbf{w}_2 &\doteq \delta\mathbf{w}_{ij}^{-\zeta}, \end{aligned}$$

where $\delta\mathbf{w}_{ij}^\zeta(j)$ is given by ξ in Eq. (3.42) and $\delta\mathbf{w}_{ij}^{-\zeta}(j)$ is given by the same equation with ζ replaced by $-\zeta$. By Lemma 16, $\{\mathbf{w} + \delta\mathbf{w}_1, \mathbf{w} + \delta\mathbf{w}_2\} \subset \mathcal{V}_K^1$. The associated contrast variations (see Eq. (2.50)) are noted

$$\begin{aligned} \Delta\tilde{C}_R^1 &\doteq \Delta\tilde{C}_R(\mathbf{w} + \delta\mathbf{w}_1, \mathbf{w}), \\ \Delta\tilde{C}_R^2 &\doteq \Delta\tilde{C}_R(\mathbf{w} + \delta\mathbf{w}_2, \mathbf{w}). \end{aligned}$$

Then, if $\zeta > 0$, either $\Delta\tilde{C}_R^1 > 0$, or $\Delta\tilde{C}_R^2 > 0$.

The proof is relegated to the Appendix 3.8.8 (p. 178). From the above theorem and the results of Section 2.3.4, one gets the following corollary.

Corollary 14 (Discriminant contrast property) *The function $\tilde{\mathcal{C}}_R$ is a discriminant contrast in the sense that $Y_i \propto S_j$ if and only if \mathbf{b}_i locally maximizes $\mathcal{C}_R(\mathbf{b}_i)$ over the unit-sphere $\mathcal{S}(K)$.*

Remark 20 (Restriction of $\mathbf{B} \in \mathcal{M}(K)$ to $\mathbf{B} \in \mathcal{O}(K)$) When proving the above results and those of Section 2.3.4, it is not always constrained that \mathbf{w} satisfy another condition than $\mathbf{w} \in \mathcal{V}_K^\lambda$. However, in order to avoid extracting twice the same source, each row vector of \mathbf{W} , i.e. the \mathbf{w}_i can always be kept orthonormal: we could e.g. constrain \mathbf{B} to belong either to $\mathcal{O}(K)$ or to $\mathcal{SO}(K)$. Hence, a natural question arises: “do the aforementioned results still hold under the additional constraint that \mathbf{B} must belong to $\mathcal{O}(K)$ or to $\mathcal{SO}(K)$?”. Clearly, if a function reaches a global maximum in a point included in $\mathbb{R}^{K \times K} \cap \mathcal{SO}(K)$, the global maximum point of this function restricted to $\mathcal{SO}(K)$ is the same point. This can be extended to the local maximum points $\mathbf{w} \in \{\mathbf{e}_1, \dots, \mathbf{e}_K\}$ of $\tilde{\mathcal{C}}_R(\mathbf{w})$ if $\mathbf{w} \in \mathcal{V}_K^1$. Indeed, since a manifold is a topological space which is locally Euclidean, for all $\mathbf{B} \in \mathcal{O}(K)$, the restriction of the neighborhood of \mathbf{B} to the manifold induced by $\mathcal{O}(K)$ is a subset of the neighborhood of \mathbf{B} in the whole $\mathbb{R}^{K \times K}$ space.

The only result that has still to be proved is that no local maximum point exists on the contrast restricted to $\mathcal{O}(K)$ for $\mathbf{w} \in \mathcal{V}_K^1 / \{\mathbf{e}_1, \dots, \mathbf{e}_K\}$. As the update of the transfer matrix must be performed through the update of the demixing matrix (in our “blind” application, remind that the only way to modify the entries of \mathbf{W} is to modify \mathbf{B}), there are actually two questions to answer. First: does the “orthogonality” restriction induce mixing maxima in the contrast function? Second, if there always exists fortunately trajectories laying on the orthogonal manifold along which we can increase the contrast if we are not in a true local maximum in the whole search space, then are these trajectories reachable by updating the demixing matrix?

Let us answer the first question. Actually, when extracting the first source, and since the last rows do not play any role in the evaluation of the contrast function, the orthogonality constraint has no impact. Assuming now that $p - 1$ sources have been extracted (that is $\mathbf{w}_i = \mathbf{e}_i$, $1 \leq i < p$); a p -th source can be recovered by constraining the $\mathbf{w}_i \mathbf{w}_j^T = \delta_{ij}$ (again, the $K - p$ last rows of \mathbf{W} do not play any role in the contrast). This implies than we are searching for \mathbf{w}_p such that the $p - 1$ first elements of \mathbf{w}_p are zero. Any update that modifies only the $K - p + 1$ last entries of \mathbf{w}_p will thus preserve the orthogonality constraint, up to an orthogonalization of the $K - p$ last rows of \mathbf{W} . Consequently, since the K -th source is trivially extracted due to the $\mathbf{W} \in \mathcal{O}(K)$ or $\mathbf{W} \in \mathcal{SO}(K)$ constraint, we can take $p < K$ and we can always proceed to the update rules $\mathbf{w}_p \leftarrow \mathbf{w}_p + \delta \mathbf{w}_1$ or $\mathbf{w}_p \leftarrow \mathbf{w}_p + \delta \mathbf{w}_2$ by taking $i, j \in \{p, \dots, K\}$. Therefore, by Theorem 18 (p. 141), the contrast can be increased if we are not in a true local maximum of the criterion (i.e. in the whole search space) because there exists $\delta \mathbf{w}$ such that $\tilde{\mathcal{C}}_R(\mathbf{w}_p + \delta \mathbf{w}) > \tilde{\mathcal{C}}_R(\mathbf{w}_p)$. On the other hand, these true local maxima

are proved by Corollary 14 to necessarily correspond to a non-mixing optima; the discriminacy property is preserved. Finally, as the last $K - p + 1$ rows have no impact on $\tilde{\mathcal{C}}_R(\mathbf{w}_p)$, it is always possible to transform them in such a way that the new transfer matrix still belongs to the orthogonal group.

Let us now turn to the second question: does there exist a small vector $\delta\mathbf{b}$ such that $\delta\mathbf{b}\mathbf{A} = \delta\mathbf{w}$ (where $\delta\mathbf{w}$ is as above)? If yes then we know that $\mathcal{C}_R(\mathbf{b}_p + \delta\mathbf{b}) > \mathcal{C}_R(\mathbf{b}_p)$ because of the above result and the relationship between $\mathcal{C}_R(\mathbf{B})$, $\tilde{\mathcal{C}}_R(\mathbf{W})$ and $\mathbf{W} = \mathbf{B}\mathbf{A}$. Clearly, the answer is positive because the columns of \mathbf{A} form an orthogonal basis of \mathbb{R}^K . But since the p first rows of \mathbf{W} are orthogonal and because of the group structure of $\mathcal{O}(K)$, the p first rows of $\mathbf{B} = \mathbf{W}\mathbf{A}^{-1}$ are also orthogonal. On the other hand one can freely orthogonalize the last $K - p$ first rows of \mathbf{B} with respect to the p first ones afterwards, in a similar way as we have done for \mathbf{W} .

Therefore, all the properties of $\mathcal{C}_R(\mathbf{b}_i)$ analyzed in $\mathbb{R}^{K \times K}$ s.t. $\|\mathbf{b}_i\| = 1$ still hold when one updates \mathbf{W} through \mathbf{B} and restricts the demixing matrix to be in the lower dimensional subset $\mathcal{O}(K) \in \mathbb{R}^{K \times K}$ at each iteration; this extends to $\mathcal{SO}(K) \in \mathbb{R}^{K \times K}$. Note however that we did not provide a mean to find either the updates $\delta\mathbf{b}$ or the matrix orthogonalization, but this is not the purpose of this remark.

3.4.2.2 Using geodesic-convexity on the hyper-sphere constraint

We assume here that \mathbf{w} is constrained to belong to the unit-sphere $\mathcal{S}(K)$. As mentioned above, the (piecewise, between two solutions) convexity of a contrast function (i.e. the concavity of its opposite) proves its discriminacy property. The definition domain $\mathcal{S}(K)$ of the function $f : \mathcal{A} \subseteq \mathcal{S}(K) \mapsto \mathbb{R}$ is no more convex, and thus the function f cannot, rigorously speaking, be convex. For instance, when $K = 2$, the domain is $\mathcal{S}(K)$, a circle with unitary radius or possibly \mathcal{V}_K^1 , the open first quadrant of this sphere. However, as shown in the preliminary section, the function is piecewise convex (in a given quadrant) when the vector \mathbf{w} is parametrized as a function of a single parameter θ , which is an implicit way to fulfill the norm constraint, and we can say that the function is piecewise convex *along a geodesic of the circle*, within a given quadrant. In other words, it could be possible to consider a kind of *geodesic convexity* (also named “g-convexity”, for short), which deals with a function in \mathbb{R}^K defined only on a lower dimensional subset (more exactly, on a sub-manifold embedded in \mathbb{R}^K [Absil et al.]), say \mathcal{A} , just as the circle (or a piece of circle) in the 2D plane. This notion has been studied by Rapcsak in [Rapcsak, 1991], and the relation between g-convexity and (non-)existence of minima/maxima has been established.

Rapcsak has investigated the geodesic convexity of a function (defined below) $f : \mathcal{A} \subset \mathcal{M} \mapsto \mathbb{R}$ where \mathcal{M} is a smooth manifold (Class 2)². The characterization of the g-convexity is based on the Lagrange function $\mathcal{L}(\mathbf{x}, \mathbf{v}(\mathbf{x}))$ where

²the details do not matter as we shall work on pieces of sphere, that correspond to manifolds in the class $C^\infty \subset C^2$.

the entries of the multiplier vector $\mathbf{v}(\mathbf{x})$ are functions. In this case, according to Rapcsak, the g-convexity consists in the positive semi-definiteness of the sum of the Hessian matrix of the objective function, noted $\mathbf{H}_{\mathbf{x}}f(\mathbf{x})$ and the so-called “second fundamental form of the manifold \mathcal{M} on a geodesic convex set \mathcal{A} ”.

Definition 22 (g-convex set) *A set \mathcal{A} is said to be g-convex if any pair of points of \mathcal{A} are joined by a geodesic belonging to \mathcal{A} .*

This definition differs from the usual one in differential geometry as we are not considering the *shortest* geodesic.

From this viewpoint, a geodesic of a sphere \mathcal{S} that links two points belonging to \mathcal{S} is any piece (i.e. not necessarily the shortest one) of the great circle that joins these points (recall that the great circles of a sphere \mathcal{S} are circles belonging to \mathcal{S} that has the same circumference as \mathcal{S} , and dividing \mathcal{S} into two equal hemispheres).

The intuitive notion is difficult to explain without many mathematical details, but it is close to the usual set convexity definition in \mathbb{R}^K . As an example, a non-convex set in \mathbb{R}^K is not g-convex in this space, but parts of a hyper-sphere may be g-convex sets: for example, in a given hyper quadrant \mathcal{Q}_p of \mathbb{R}^K , the corresponding piece of sphere

$$\mathcal{S}(K) \cap \mathcal{Q}_p$$

is g-convex. This would not be the case if the “cutting edges” were more complicated.

Based on the concept of g-convex set, we can define the geodesic convexity of a function.

Definition 23 (g-convex function) *Let $\mathcal{A} \subset \mathcal{M}$ be a g-convex set. Then, it is said that a function $f : \mathcal{A} \mapsto \mathbb{R}$ is g-convex if its restriction to all geodesic arcs belonging to \mathcal{A} are convex in the arc length parameter.*

It results from the definition that the following inequalities hold for every geodesic $g(s)$, $s \in [0, b]$, joining two arbitrary points $m_1, m_2 \in \mathcal{A}$:

$$f(g(tb)) \leq (1-t)f(g(0)) + tf(g(b)), \quad 0 \leq t \leq 1,$$

where $g(0) = m_1$, $g(b) = m_2$ and s is the arc length parameter. The above definition is close to the usual definition of convex function.

We have the following theorem.

Theorem 19 (Rapcsak [Rapcsak, 1991]) *Let $\mathcal{A} \subseteq \mathcal{M}$ be an open g-convex set where \mathcal{M} is a connected manifold defined by $K - n$ one-dimensional constraints $\gamma_j(\mathbf{x})$ as*

$$\mathcal{M} \doteq \{\mathbf{x} | \gamma_j(\mathbf{x}) = 0, j = 1, \dots, K - n, \mathbf{x} \in \mathbb{R}^K\} .$$

Let $f : \mathcal{A} \mapsto \mathbb{R}$ be a twice continuously differentiable function and the $\gamma_j(\mathbf{x})$ be linearly independent. Then f is g-convex on \mathcal{A} if and only if the matrix

$\mathbf{H}_{\mathbf{x}}^g \mathcal{L}(\mathbf{x}, \mathbf{v}(\mathbf{x}))|_{T_{\mathcal{M}}}$ is positive semi-definite at every point $\mathbf{x} \in \mathcal{A}$. The above matrix is defined by

$$\mathbf{H}_{\mathbf{x}}^g \mathcal{L}(\mathbf{x}, \mathbf{v}(\mathbf{x}))|_{T_{\mathcal{M}}} \doteq \left[\mathbf{H}_{\mathbf{x}} f(\mathbf{x}) - \sum_{j=1}^{K-n} v_j(\mathbf{x}) \mathbf{H}_{\mathbf{x}} \gamma_j(\mathbf{x}) \right]_{T_{\mathcal{M}}} , \quad (3.44)$$

where $\mathbf{H}_{\mathbf{x}} f(\mathbf{x})$, $\mathbf{H}_{\mathbf{x}} \gamma_j(\mathbf{x})$ are the Hessian of $f(\mathbf{x})$ and $\gamma_j(\mathbf{x})$, respectively, and $v_j(\mathbf{x})$ are the functional Lagrange multipliers, defined in a vector formed by

$$\mathbf{v}^T(\mathbf{x}) \doteq \nabla f(\mathbf{x}) \nabla \gamma^T(\mathbf{x}) [\nabla \gamma(\mathbf{x}) \nabla \gamma^T(\mathbf{x})]^{-1} ,$$

with

$$\nabla \gamma(\mathbf{x}) \doteq \begin{pmatrix} \nabla \gamma_1(\mathbf{x}) \\ \vdots \\ \nabla \gamma_{K-n}(\mathbf{x}) \end{pmatrix} .$$

In the above theorem, the subscript $T_{\mathcal{M}}$ denotes the projection operator onto the tangent space of \mathcal{M} at \mathbf{x} . It can be shown [Luenberger, 1973] that

$$\mathbf{H}_{\mathbf{x}}^g \mathcal{L}(\mathbf{x}, \mathbf{v}(\mathbf{x}))|_{T_{\mathcal{M}}} = \mathbf{P}^T \mathbf{H}_{\mathbf{x}}^g \mathcal{L}(\mathbf{x}, \mathbf{v}(\mathbf{x})) \mathbf{P} , \quad (3.45)$$

where \mathbf{P} is the projection matrix defined by

$$\mathbf{P} \doteq \mathbf{I}_K - \nabla \gamma^T(\mathbf{x}) [\nabla \gamma(\mathbf{x}) \nabla \gamma^T(\mathbf{x})]^{-1} \nabla \gamma(\mathbf{x}) .$$

The g-convexity of a contrast function f is related to the possible discriminacy property of f by Corollary 3.1 of [Rapcsak, 1991]. Basically, this corollary says that if a function $f : \mathcal{A} \subset \mathcal{M} \mapsto \mathbb{R}$ is a continuously differentiable g-convex function, and \mathcal{A} is an open g-convex set, then every stationary point of f is a global minimum; in other words, the local maxima are necessarily attained at a boundary of \mathcal{A} (where the derivative of f might not exist). Hence, if a piecewise-differentiable contrast function is g-convex in “each piece” (shortly, if f is *piecewise g-convex*) between the solution points, it is a discriminant contrast function.

Let us now specialize to our range-based functional. The main problem is again the fact that $f(\mathbf{w}) \doteq -R(\mathbf{w}\mathbf{S})$ is not everywhere differentiable: it does not fulfill an essential assumption of Theorem 19 as it is only differentiable with respect to the w_i variable at points where $w_i \neq 0$, because of the absolute value in Eq. (2.40).

Therefore, we shall *partition* the search space into open subsets in which $f(\mathbf{w})$ is infinitely differentiable. More precisely, the idea is to prove the g-convexity property in restrictions of $f(\mathbf{w})$ to subsets of $\mathcal{S}(K)$. As an illustration, for $K = 3$, we prove that the function $f(\mathbf{w})$, $\mathbf{w} \in \mathcal{S}(3)$ is g-convex within each of the hyper-quadrants of the sphere (quadrant boundaries excluded); as the function is everywhere differentiable in these subsets, this means, by Rapcsak’s Corollary 3.1, that vectors $\mathbf{w} \in \mathcal{S}(3)$ of the form $w_1 w_2 w_3 \neq 0$ cannot be local

maximum points of f (if there exists a stationary point in these subsets, it is a global minimum). Then we focus on the restrictions of f to $\{\mathbf{w} \in \mathcal{S}(3), w_i = 0\}$ for $i \in \{1, 2, 3\}$. We prove then the piecewise g-convexity of these restricted functions in the corresponding circle $w_j^2 + w_k^2 = 1$, where $j, k \in \{1, 2, 3\} \setminus \{i\}$ and $j \neq k$ (i.e. the g-convexity of these restrictions in each quadrant of these circles, again boundaries excluded); in these subsets, the restricted function is continuously differentiable with respect to the variables not corresponding to the i -th entry of \mathbf{w} . Finally, applying again Rapcsak's Corrolary 3.1, the above procedure proves that local maxima of $f(\mathbf{w})$, $\mathbf{w} \in \mathcal{S}(3)$ can only be attained at vectors of the form $w_i = w_j = 0$, $|w_k| = 1$ where (i, j, k) is a permutation of $(1, 2, 3)$.

Note that looking for the local maxima of f in each of these subsets is equivalent to looking for these maxima in the whole search space because the collection of these subsets covers $\mathcal{S}(K)$. In the general K -dimensional space, the search set $\mathcal{S}(K)$ can be decomposed as

$$\mathcal{S}(K) = \bigcup_{k=1}^K \{\mathbf{w} \in \mathcal{S}(K) : \#[I(\mathbf{w})] = k\} , \quad (3.46)$$

where $I(\mathbf{w})$ is defined in Eq. (1.96): $\mathcal{S}(K)$ is written as the union of the unit-norm vectors having k non-zero entries ($1 < k \leq K$). As an example, the K -entries basis vectors belong to $\{\mathbf{w} \in \mathcal{S}(K) : \#[I(\mathbf{w})] = 1\}$ and the vectors of the form $[\sin \alpha, \cos \alpha, 0, \dots, 0]$, $[\sin \alpha, 0, \cos \alpha, 0, \dots, 0]$ (etc) belong to the set $\{\mathbf{w} \in \mathcal{S}(K) : \#[I(\mathbf{w})] = 2\}$ (prevent if $\alpha = k\pi/2$ for some $k \in \mathbb{Z}$ since in this case, $\#[I(\mathbf{w})] = 1$), and so on.

We shall analyze the geodesic convexity in each of these subsets: we first focus on the $k = K$ subset before investigating the $k < K$ ones. The subset $\{\mathbf{w} \in \mathbb{R}^K : \#[I(\mathbf{w})] = K\}$ is not connected because the vectors with a positive i -th entry ($i \in I(\mathbf{w})$) cannot be joined to vectors with a negative i -th entry without crossing the hyperplane $\{\mathbf{w} : w_i = 0\}$ that does not belong to this set. Nevertheless, the subset can be rewritten as a union of 2^K connected subsets as follows:

$$\{\mathbf{w} \in \mathcal{S}(K) : \#[I(\mathbf{w})] = K\} = \bigcup_{n=1}^{2^K} \{\mathbf{w} \in \mathcal{S}(K) : \#[I(\mathbf{w})] = K, \text{sign}(\mathbf{w}) = \mathbf{1}_{\pm}^{K,n}\} , \quad (3.47)$$

where $\mathbf{1}_{\pm}^{k,n}$ denotes any k -entries vector with entries in $\{-1, 1\}$ (the 2^K superscript results from the choice of the sign of the w_i , i.e. to the number of different vectors $\mathbf{1}_{\pm}^{K,n}$). The above subsets are actually the hyper-quadrants of $\mathcal{S}(K)$ (boundaries excluded, i.e. open): they are located on a same side of the hyperplanes $\{\mathbf{w} \in \mathbb{R}^K : w_1 = 0\}, \dots, \{\mathbf{w} \in \mathbb{R}^K : w_K = 0\}$, and not on these planes as $\#[I(\mathbf{w})] = K$. They are thus open g-convex connected sets.

Example 20 For a given k , if $\alpha \in]k\pi/2, (k+1)\pi/2[$, $[\sin \alpha, \cos \alpha] \in \{\mathbf{w} \in \mathcal{S}(2) : \#[I(\mathbf{w})] = 2, \text{sign}(\mathbf{w}) = \mathbf{1}_{\pm}^{2,n}\}$ for some n and if $\alpha \in](k+1)\pi/2, (k+2)\pi/2[$, this vector belongs to $\{\mathbf{w} \in \mathcal{S}(2) : \#[I(\mathbf{w})] = 2, \text{sign}(\mathbf{w}) = \mathbf{1}_{\pm}^{2,n'}\}$ for $n' \neq n$.

To check if $f(\mathbf{w})$ is piecewise g-convex in the set $\{\mathbf{w} \in \mathcal{S}(K) : \#[I(\mathbf{w})] = K\}$, we are led to compute $\mathbf{H}_{\mathbf{w}}^g \mathcal{L}(\mathbf{w}, \mathbf{v}(\mathbf{w}))$ in each of the connected subsets given in the

right-hand side of (3.47), part of the smooth manifold $\mathcal{S}(K)$, on which $f(\mathbf{w})$ is infinitely differentiable. Hence Theorem 19 holds. The constraint ($K - n = 1$) is $\gamma(\mathbf{w}) \equiv \|\mathbf{w}\|^2 - 1 = 0$. This yields $f(\mathbf{w}) = -\sum_{l=1}^K |w_l| R(S_l)$, $|w_i| > 0$ for all i and

$$\nabla f(\mathbf{w}) = -[\text{sign}(w_1)R(S_1), \dots, \text{sign}(w_K)R(S_K)] \quad (3.48)$$

$$\mathbf{H}_{\mathbf{w}} f(\mathbf{w}) = \mathbf{0} . \quad (3.49)$$

On the other hand, $\nabla \gamma(\mathbf{w}) = 2\mathbf{w}$, and the functional multiplier is

$$v(\mathbf{w}) = -2[\text{sign}(w_1)R(S_1), \dots, \text{sign}(w_K)R(S_K)]\mathbf{w}^T \frac{1}{4\|\mathbf{w}\|^2} = -\frac{R(\mathbf{w}S)}{2\|\mathbf{w}\|^2} .$$

In the above relations, we have used $\text{sign}(x) = \pm 1$ depending if $x \gtrless 0$ and may be either $+1$ or -1 if $x = 0$. In what follows, $\text{sign}(\mathbf{w})$ has to be understood as the vector having as i -th component the sign of w_k . Hence, as $\mathbf{H}_{\mathbf{w}} \gamma(\mathbf{w}) = 2\mathbf{I}_K$, we find

$$\begin{aligned} \mathbf{H}_{\mathbf{w}}^g \mathcal{L}(\mathbf{w}, v(\mathbf{w}))|_{T_{\mathcal{M}}} &= \mathbf{P}^T (\mathbf{0} - 2v(\mathbf{w})\mathbf{I}_K) \mathbf{P} \\ &= -\frac{R(\mathbf{w}S)}{\|\mathbf{w}\|^2} \mathbf{P}^T \mathbf{P} , \end{aligned} \quad (3.50)$$

which proves that the above Hessian is negative semi-definite because $\mathbf{P}^T \mathbf{P} \succeq 0$. Consequently, the function $f(\mathbf{w}) = -R(\mathbf{w}S)$ is g-convex at any point of the subsets given in (3.47) (i.e. into an open hyper-quadrant of $\mathcal{S}(K)$). Consequently, by Rapcsak's Corollary 3.1, there is no local maximum point of f in these subsets.

Let us now analyze $f(\mathbf{w})$ inside the other subsets $\{\mathbf{w} \in \mathcal{S}(K) : \#[I(\mathbf{w})] = k\}$ for some $1 < k < K$ (the $k = K$ case has been analyzed right above, and the $k = 1$ case corresponds to non-mixing points, in which we are not interested in this chapter). Consider the further decomposition of this set as

$$\cup_{m=1}^{K'} \cup_{n=1}^{K''} \{\mathbf{w} \in \mathcal{S}(K) : \#[I_{m,n}] = k, I(\mathbf{w}) = I_{m,n}, \widehat{\text{sign}}(\mathbf{w}) = \mathbf{1}_{\pm}^{k,n}\} , \quad (3.51)$$

with $K' \doteq \binom{K}{k}$ is the number of possibilities to choose the index vectors $I_{m,n}$ such that $\#[I_{m,n}] = k$ and for each of them, there are still $K'' \doteq 2^k$ ways to choose $\mathbf{1}_{\pm}^{k,n}$. In the above equation, $I_{i,n} = I_{j,m}$ iff $i = j$ and the elements of $I_{m,n} \subset \{1, \dots, K\}$ are all distinct. The $\widehat{\text{sign}}(\mathbf{w})$ term is called *the subvector of sign(\mathbf{w}) with respect to $I(\mathbf{w})$* :

Definition 24 (subvectors) *The subvector $\widehat{\mathbf{w}}$ of $\mathbf{w} \in \mathbb{R}^K$ with respect to a set of distinct indexes $I \subseteq \{1, \dots, K\}$ is defined as the vector having its i -th entry, noted $\widehat{w}_i = \widehat{\mathbf{w}}(i)$, set equal to $w_{i'} = \mathbf{w}(i')$, where i' denotes the i -th smallest component of I (or equivalently, the i -th component of a permuted version of I in which the components are sorted by increasing values). Its dimension corresponds to the number of elements in I : $\widehat{\mathbf{w}} \in \mathbb{R}^{\#[I]}$.*

A subvector is thus a lower-dimensional “down-sampled” version of the original vector built according to the indexes (sorted by increasing order) of the non-zero entries of \mathbf{w} . It results from the definition that a subvector $\widehat{\mathbf{w}}$ of \mathbf{w} with respect to $I(\mathbf{w})$ has no zero-valued component.

Example 21 If $\mathbf{w} = [0.3, 0, 0.78, .32, 0, .45]$ and $\mathbf{v} = [.21, -.1, -.2, .4, .2, 0]$, their subvectors with respect to $I(\mathbf{w}) = \{1, 4, 3, 6\}$ reduce respectively to $\widehat{\mathbf{w}} = [0.3, 0.78, .32, .45]$ and $\widehat{\mathbf{v}} = [.21, -.2, .4, 0]$.

Note that the order of the elements of a set of indexes does not matter (e.g. $\{1, 4, 3, 6\} = \{1, 3, 4, 6\}$). Further, it is obvious that $\widehat{\text{sign}(\mathbf{w})} = \text{sign}(\widehat{\mathbf{w}})$ when the reference index set used for building these subvectors is $I(\mathbf{w})$ in both cases.

Example 22 Let n be such that $\mathbf{1}_{\pm}^{2,n} = [1, -1]$ and m corresponding to $I_{m,n} = \{1, 3\}$. Consider the vector $[\sin \alpha, 0, \cos \alpha, 0, \dots, 0] \in \mathcal{S}(K)$. Then, this vector belongs to $\{\mathbf{w} \in \mathcal{S}(K) : \#[I_{m,n}] = 2, I(\mathbf{w}) = I_{m,n}, \widehat{\text{sign}(\mathbf{w})} = \mathbf{1}_{\pm}^{2,n}\}$ iff α belongs to the second quadrant of the unit circle, boundaries excluded.

Each of the subsets of the right-hand side of (3.51) is nothing but the set of K -entries unit-norm vectors having their k non-zero entries at the same places (they are given by the set of indexes $I(\mathbf{w}) = I_{m,n}$) and in which each entry have also the same sign (fixed by the vector $\mathbf{1}_{\pm}^{2,n}$ via the index n). These subsets correspond to some hyper-quadrants of the k -dimensional hyper-sphere. Indeed, the restriction of $\mathcal{S}(K)$ to a lower-dimensional linear subspace of the form $\{\mathbf{w} \in \mathcal{S}(K) : \#[I_{m,n}] = k < K, I(\mathbf{w}) = I_{m,n}, \widehat{\text{sign}(\mathbf{w})} = \mathbf{1}_{\pm}^{k,n}\}$ also forms a sub-manifold, say \mathcal{M}' which corresponds to a piece of a lower-dimensional hyper-sphere embedded in \mathbb{R}^k . For example the sphere $\mathcal{S}(3)$ restricted to the space spanned by the elements of the set $\{\mathbf{w} \in \mathbb{R}^3 : w_3 = 0\}$ is the *great circle* of $\mathcal{S}(3)$ given by $w_1^2 + w_2^2 = 1$; it can be divided in four quadrants in which the sign and the number of the non-zero entries of the vectors located in any of these quadrants are kept constant. These pieces of sphere are open g-convex sets. Consequently, as the range functional is infinitely differentiable within such hyper-quadrants, we can apply exactly the same reasoning as in the above $k = K$ case to show that the criterion is (piecewise) g-convex on each of the \mathcal{M}' : it suffices to consider $\widehat{\mathbf{w}}$ instead of \mathbf{w} , and k instead of K .

Applying successively this reasoning, it is shown than whatever is the subset $\{\mathbf{w} \in \mathcal{S}(K) : \#[I(\mathbf{w})] = k\}$, $1 \leq k \leq K$, the function is g-convex within sub-g-convex sets forming a partition of the above set. Therefore, by Rapcsak's Corollary 3.1, the local maxima can only be reached on the boundaries (no “inner maxima” inside the subspace): if $k > 1$, maximizing locally $f(\mathbf{w})$ in a given $\{\mathbf{w} \in \mathcal{S}(K) : \#[I(\mathbf{w})] = k\}$ would produce a new \mathbf{w} satisfying necessarily $\#[I(\mathbf{w})] \leq k - 1$ and iterating this result on successive sub-manifolds shows that all the local maximum points of $f(\mathbf{w})$ in the whole search space $\mathcal{S}(K)$ are the basis vectors ($\#[I(\mathbf{w})] = 1$).

3.4.2.3 Using the Hessian of the penalized output range

In this section, we focus on the unconstrained output range criterion $-\tilde{\mathcal{C}}_R(\mathbf{w}) = R(\mathbf{w}\mathbf{S}/\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]})$; the vector $\mathbf{w} \in \mathbb{R}^K$ is no longer constrained to belong to $\mathcal{S}(K)$. We shall prove that none of the stationary points of the criterion can correspond to a local minimum (we consider the opposite of the contrast function) if two entries of \mathbf{w} are non zero. Remind that

$$R\left(\mathbf{w}\mathbf{S}/\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]}\right) = \frac{\sum_{l=1}^K |w_l|R(\mathbf{S}_l)}{\|\mathbf{w}\|}. \quad (3.52)$$

From this, we have the following lemma, proved in Section 3.8.9 (p. 179). Again, because of the absolute values in the criterion, only the derivatives at points where all entries of \mathbf{w} are non-zero exist. Remind that the mixing extremum points are such that $\#[I(\mathbf{w})] \geq 2$.

Lemma 17 (Derivatives of $R\left(\mathbf{w}\mathbf{S}/\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]}\right)$) *The first derivative of the criterion $R\left(\mathbf{w}\mathbf{S}/\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]}\right)$ at point \mathbf{w} with respect to the variables corresponding to non-zero entries in \mathbf{w} is:*

$$\frac{\partial R\left(\mathbf{w}\mathbf{S}/\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]}\right)}{\partial w_k} = \frac{\text{sign}(w_k)R(\mathbf{S}_k)\|\mathbf{w}\|^2 - w_k \sum_{l=1}^K |w_l|R(\mathbf{S}_l)}{\|\mathbf{w}\|^3}, \quad k \in I(\mathbf{w}). \quad (3.53)$$

The second derivatives at stationary points are

$$\frac{\partial^2 R\left(\mathbf{w}\mathbf{S}/\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]}\right)}{\partial w_k^2} = \frac{|w_k|R(\mathbf{S}_k)}{\|\mathbf{w}\|^3} \left(1 - \frac{\|\mathbf{w}\|^2}{w_k^2}\right), \quad k \in I(\mathbf{w}), \quad (3.54)$$

$$\frac{\partial^2 R\left(\mathbf{w}\mathbf{S}/\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]}\right)}{\partial w_i \partial w_j} = \frac{w_i w_j}{\|\mathbf{w}\|^5} \sum_{l=1}^K |w_l|R(\mathbf{S}_l), \quad i \neq j, (i, j) \in I(\mathbf{w}) \quad (3.55)$$

From the above lemma, one gets the announced result (the proof will be given within the core of the text because of its simplicity and its shortness).

Lemma 18 *For any mixing stationary point (i.e. satisfying $\|\mathbf{w}\| \neq \|\mathbf{w}\|_\infty$ or, equivalently, $\#[I(\mathbf{w})] \geq 2$), the criterion $R\left(\mathbf{w}\mathbf{S}/\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]}\right)$ does not have a local minimum (and the associated contrast $\tilde{\mathcal{C}}_R(\mathbf{w})$ does not have a local maximum).*

Proof:

In order to prove this lemma, we have to show that the Hessian evaluated at any mixing point cannot be positive definite (remind that at these points \mathbf{w} , there exists a pair i, j of indexes satisfying $i \in I(\mathbf{w}), j \in I(\mathbf{w})$ such that the derivatives of the criterion at point \mathbf{w} with respect to the variables corresponding to the i, j

entries \mathbf{w} exist). We shall show that it is non-positive definite. From Lemma 17, we are led to compute

$$\begin{aligned} \frac{1}{\|\mathbf{w}\|^3} & [\epsilon \ \eta] \begin{bmatrix} \frac{R(\mathbf{S}_i)}{|w_i|}(w_i^2 - \|\mathbf{w}\|^2) & \frac{w_i w_j}{\|\mathbf{w}\|^2} \sum_{l=1}^K |w_l| R(\mathbf{S}_l) \\ \frac{w_i w_j}{\|\mathbf{w}\|^2} \sum_{l=1}^K |w_l| R(\mathbf{S}_l) & \frac{R(\mathbf{S}_j)}{|w_j|}(w_j^2 - \|\mathbf{w}\|^2) \end{bmatrix} \begin{bmatrix} \epsilon \\ \eta \end{bmatrix} \\ & = \underbrace{\frac{\sum_{l=1}^K |w_l| R(\mathbf{S}_l)}{\|\mathbf{w}\|^5} [\epsilon \ \eta]}_{\doteq \alpha \geq 0} \begin{bmatrix} w_i^2 - \|\mathbf{w}\|^2 & w_i w_j \\ w_j w_i & w_j^2 - \|\mathbf{w}\|^2 \end{bmatrix} \begin{bmatrix} \epsilon \\ \eta \end{bmatrix}, \quad (3.56) \end{aligned}$$

where we have used the stationary point condition $\partial R \left(\mathbf{wS} / \sqrt{\text{Var}[\mathbf{wS}]} \right) / \partial w_i = 0$, i.e. from Eq. (3.53):

$$R(\mathbf{S}_i) / |w_i| = \frac{1}{\|\mathbf{w}\|^2} \sum_{l=1}^K |w_l| R(\mathbf{S}_l) . \quad (3.57)$$

Hence,

$$\begin{aligned} \alpha(\epsilon^2 w_i^2 - \epsilon^2 \|\mathbf{w}\|^2 + 2\epsilon\eta w_i w_j + \eta^2 w_j^2 - \eta^2 \|\mathbf{w}\|^2) & = -\alpha(\epsilon^2 + \eta^2) \sum_{l=1, l \neq \{i,j\}}^K w_l^2 \\ & \quad -\alpha(\epsilon w_j - \eta w_i)^2 \\ & \leq 0 , \quad (3.58) \end{aligned}$$

whatever the vector $[\epsilon, \eta]$.

□

The above lemma says that if a mixing stationary point of the criterion $R \left(\mathbf{wS} / \sqrt{\text{Var}[\mathbf{wS}]} \right)$ exists, it cannot be a local minimum.

3.4.3 Simultaneous approach

As the simultaneous approach to BSS is a particular case of the partial simultaneous approach with $P = K$, one could directly turn to the discriminacy property of the range-based partial contrast $\mathcal{C}_R(\mathbf{B})$, $\mathbf{B} \in \mathbb{R}^{P \times K}$. Nevertheless, the proof of the discriminacy is more involved for the partial contrast than for the simultaneous one, so that we provide the proof for $P = K$ first, as an introduction. This work has been published in [Pham and Vrins, 2006].

The contrast property of $\mathcal{C}_R : \mathbf{B} \in \mathbb{R}^{K \times K} \mapsto \mathbb{R}$ means that $\tilde{\mathcal{C}}_R(\mathbf{W})$ attains its global maximum at and only at matrices $\mathbf{W} \sim \mathbf{I}_K$, $\mathbf{W} = \mathbf{BA}$. In the remaining part of this subsection, it is shown that there exists no other local maximum of this criterion.

The function $\tilde{\mathcal{C}}_R$ is not everywhere differentiable on $\mathcal{M}(K)$, due to the absolute value in (2.42). To overcome this difficulty, we introduce the subsets $\mathcal{M}_I(K)$ of $\mathcal{M}(K)$, indexed by subsets I of $\{1, \dots, K\} \times \{1, \dots, K\}$, defined by

$$\mathcal{M}_I(K) = \{\mathbf{W} \in \mathcal{M}(K) : W_{ij} \neq 0 \text{ if and only if } (i, j) \in I\}. \quad (3.59)$$

Example 23 For example, the top matrices given in Ex. 1 (p. 8) belong to (from left to right) $\mathcal{M}_{I1}(3)$ and $\mathcal{M}_{I2}(3)$, respectively, iff $I1 = \{(1, 1), (2, 3), (3, 2)\}$ and $I2 = \{(1, 2), (2, 2), (2, 3), (3, 1)\}$.

Due to the non-singularity condition $\mathbf{W} \in \mathcal{M}(K)$, a subset $\mathcal{M}_I(K)$ may be empty for a particular I . For example, if I is a subset of $\{1, \dots, K\} \times \{1, \dots, K\}$ such that its i -th section $I_i \doteq \{j \in \{1, \dots, K\}, (i, j) \in I\}$ is empty for some $i \in \{1, \dots, K\}$, then any matrix \mathbf{W} such that $W_{ij} = 0$ if $(i, j) \notin I$ would be singular, hence $\mathcal{M}_I(K)$ is empty³. Thus we shall restrict ourselves to the collection \mathcal{I} of distinct subsets I of $\{1, \dots, K\} \times \{1, \dots, K\}$ such that $\mathcal{M}_I(K)$ is not empty. Then the subsets $\mathcal{M}_I(K)$, $I \in \mathcal{I}$, form a partition of $\mathcal{M}(K)$, since they are clearly disjoint and their union covers $\mathcal{M}(K)$. Therefore any local maximum point of $\tilde{\mathcal{C}}_R$ would belong to some $\mathcal{M}_I(K)$ with $I \in \mathcal{I}$ and is necessarily a local maximum point of the restriction of $\tilde{\mathcal{C}}_R$ on $\mathcal{M}_I(K)$.

The key point is that the restriction of $\tilde{\mathcal{C}}_R$ to $\mathcal{M}_I(K)$, $I \in \mathcal{I}$, is infinitely differentiable as a function of the nonzero elements of its matrix argument in $\mathcal{M}_I(K)$. Thus, one may look at the first and second derivatives of the restriction of $\tilde{\mathcal{C}}_R$ to $\mathcal{M}_I(K)$ to identify its local maximum points.

Lemma 19 For $I \in \mathcal{I}$, the restriction of $\tilde{\mathcal{C}}_R$ to $\mathcal{M}_I(K)$ admits the first and second partial derivatives

$$\begin{aligned} \frac{\partial \tilde{\mathcal{C}}_R(\mathbf{W})}{\partial W_{ij}} &= W^{ij} - \frac{\text{sign}(W_{ij})R(\mathbf{S}_j)}{\sum_{l=1}^K |W_{il}|R(\mathbf{S}_l)}, & (i, j) \in I \\ \frac{\partial^2 \tilde{\mathcal{C}}_R(\mathbf{W})}{\partial W_{ij} \partial W_{kl}} &= -W^{kj}W^{il}, & (i, j), (k, l) \in I, k \neq i, \\ \frac{\partial^2 \tilde{\mathcal{C}}_R(\mathbf{W})}{\partial W_{ij} \partial W_{il}} &= \frac{\text{sign}(W_{ij})\text{sign}(W_{il})R(\mathbf{S}_j)R(\mathbf{S}_l)}{[\sum_{k=1}^K |W_{ik}|R(\mathbf{S}_k)]^2} - W^{ij}W^{il}, & (i, j), (i, l) \in I, \end{aligned} \quad (3.60)$$

where W^{ij} denote the (j, i) element of \mathbf{W}^{-1} : $W^{ij} \doteq [\mathbf{W}^{-1}]_{ji}$.

The above lemma, proved in Section 3.8.10 (p. 180) allows one to characterize the stationary points of the restriction of $\tilde{\mathcal{C}}_R$ to $\mathcal{M}_I(K)$, by setting its derivative to zero, yielding the stationary point for the range-based simultaneous BSS contrast:

$$W^{ij} = \frac{\text{sign}(W_{ij})R(\mathbf{S}_j)}{\sum_{l=1}^K |W_{il}|R(\mathbf{S}_l)}, \quad (i, j) \in I. \quad (3.61)$$

³ A simple $K = 2$ example is $I = \{(2, 1), (2, 2)\}$: the first row of the matrices belonging to this set $\mathcal{M}_I(K)$ are zero and consequently, these matrices are singular.

Thus one gets the following corollary, proved in Section 3.8.11 (p. 181).

Corollary 15 *Let $I \in \mathcal{I}$, then for any $\mathbf{W} \in \mathcal{M}_I(K)$ which is a stationary point of the restriction of $\tilde{\mathcal{C}}_R$ on $\mathcal{M}_I(K)$:*

$$\{(i, j) \in \{1, \dots, K\}^2 : W^{ij} \neq 0\} \supseteq I \quad (3.62)$$

and

$$\frac{\partial^2 \tilde{\mathcal{C}}_R(\mathbf{W})}{\partial W_{ij} \partial W_{il}} = 0, \quad (i, j) \in I, (i, l) \in I. \quad (3.63)$$

The above corollary is the key point for proving the following lemma (see the proof in Section 3.8.12, p. 181).

Lemma 20 *Let $I \in \mathcal{I}$ be such that there exists a pair of indices i, j in $\{1, \dots, K\}$ for which the i -th and the j -th sections of I are not disjoint (the i -th section of I is the set $\{k \in \{1, \dots, K\} : (i, k) \in I\}$). Then the restriction of $\tilde{\mathcal{C}}_R$ in $\mathcal{M}_I(K)$ does not have a local maximum point.*

Lemma 20 allows one to eliminate subsets I in \mathcal{I} for which the restriction of $\tilde{\mathcal{C}}_R$ in $\mathcal{M}_I(K)$ does not have a local maximum point. The only subset left is the one such that its i -th sections reduce to a single point, for all $i = 1, \dots, K$ and are disjoint (one element per row and per column). In other words, only matrices $\mathbf{W} \in \mathcal{M}(K)$ could yield a local maximum point of $\tilde{\mathcal{C}}_R(\mathbf{W})$, and those matrices have been proved to yield indeed, a global maximum point (see Section 2.3.3). This yields the discriminacy property of $\tilde{\mathcal{C}}_R$.

Corollary 16 (Discriminacy of $\mathcal{C}_R(\mathbf{B})$, simultaneous approach) *The only local maximum points of $\mathcal{C}_R(\mathbf{B})$, $\mathbf{B} \in \mathcal{M}(K)$ are the matrices $\mathbf{W} \sim \mathbf{I}$. (They are also the global minimum points.)*

3.4.4 Partial approach

In this section, we extend the discriminacy result of the simultaneous method to the partial separation approach. The results appeared in [Vrins and Pham, 2007].

In order to analyze the possible existence of mixing maxima of $\mathcal{C}_R(\mathbf{B})$ (i.e. of $\tilde{\mathcal{C}}_R(\mathbf{W})$, with $\mathbf{W} = \mathbf{B}\mathbf{A}$), we shall first compute the first two derivatives of $\log |\det(\mathbf{WW}^T)|$ with respect to the entries W_{ij} of \mathbf{W} , as for the simultaneous contrast; they are provided in the following lemma (some useful mathematical relations involved in the proof, relegated in the Chapter appendix in Section 3.8.13, p. 181 are taken from [Bernstein, 1954, Graybill, 1983, Harville, 1997, Petersen and Pedersen, 2005]).

Lemma 21 *Let $\mathbf{W} \in \mathcal{M}^{P \times K}$ and denote by $\mathbf{W}^+ \doteq \mathbf{W}^T(\mathbf{WW}^T)^{-1}$ its pseudo-inverse. Then*

$$\frac{\partial \log |\det(\mathbf{WW}^T)|}{\partial W_{ij}} = 2[(\mathbf{W}^+)^T]_{ij} = 2[\mathbf{W}^+]_{ji},$$

and

$$\frac{\partial^2 \log |\det(\mathbf{W}\mathbf{W}^T)|}{\partial W_{kl} \partial W_{ij}} = 2 \left\{ [(\mathbf{W}\mathbf{W}^T)^{-1}]_{ki} (\delta_{jl} - [\mathbf{W}^+ \mathbf{W}]_{lj}) - [\mathbf{W}^+]_{li} [\mathbf{W}^+]_{jk} \right\},$$

where δ_{jl} is the Kronecker delta. Remind that if $P = K$, $\mathbf{W}^+ = \mathbf{W}^{-1}$ and $[\mathbf{W}^+ \mathbf{W}]_{lj} = \delta_{lj}$.

Let us now use the above results for computing the first and second order derivatives of $\tilde{C}_R(\mathbf{W})$. As in the simultaneous case, we are facing the problem that \tilde{C}_R is not everywhere differentiable on $\mathcal{M}^{P \times K}$. To overcome this difficulty, we use a similar trick as for the S-BSS contrast, and, similarly to $\mathcal{M}_I(K)$ we introduce the subsets $\mathcal{M}_I^{P \times K}$ of $\mathcal{M}^{P \times K}$, indexed by subsets I of $\mathbb{Z}^{P \times K} \doteq \{1, \dots, P\} \times \{1, \dots, K\}$, defined by

$$\mathcal{M}_I^{P \times K} \doteq \{\mathbf{W} \in \mathcal{M}^{P \times K} : W_{ij} \neq 0 \text{ if and only if } (i, j) \in I\}, \quad (3.64)$$

which obviously satisfies $\mathcal{M}_I^{K \times K} = \mathcal{M}_I(K)$.

Example 24 For example, the bottom matrices given in Ex. 1 (p. 8) belong to (from left to right) $\mathcal{M}_{I_1}^{3 \times 4}$ and $\mathcal{M}_{I_2}^{2 \times 3}$, respectively, iff $I_1 = \{(1, 1), (2, 4), (3, 2)\}$ and $I_2 = \{(1, 1), (2, 2), (2, 3)\}$.

Again, due to the $\mathbf{W} \in \mathcal{M}^{P \times K}$ restriction, a subset $\mathcal{M}_I^{P \times K}$ may be empty for particular I . For example, if I is a subset of $\mathbb{Z}^{P \times K}$ such that its i -th section $I_i \doteq \{j \in \{1, \dots, K\}, (i, j) \in I\}$ is empty for some $i \in \{1, \dots, P\}$, then any matrix \mathbf{W} such that $W_{ij} = 0$ if $(i, j) \notin I$ (including all matrices \mathbf{W} s.t. $W_{ij} \neq 0$ if and only if $(i, j) \in I$) satisfy $\text{rank}(\mathbf{W}) < P$. Then, $\mathcal{M}_I^{P \times K}$ is empty because $\mathcal{M}_I^{P \times K} \subset \mathcal{M}^{P \times K}$ by definition.

Thus we shall restrict ourselves to the collection \mathcal{I} of distinct subsets I of $\mathbb{Z}^{P \times K}$ such that $\mathcal{M}_I^{P \times K}$ is not empty. Then the subsets $\mathcal{M}_I^{P \times K}$, $I \in \mathcal{I}$, form a partition of $\mathcal{M}^{P \times K}$, since they are clearly disjoint and their union equals $\mathcal{M}^{P \times K}$. Therefore any local maximum point of $\tilde{C}_R(\mathbf{W})$ ($\mathbf{W} \in \mathcal{M}^{P \times K}$) would belong to some $\mathcal{M}_I^{P \times K}$ with $I \in \mathcal{I}$ and is necessarily a local maximum point of the restriction of \tilde{C}_R on $\mathcal{M}_I^{P \times K}$.

The key point is that, here again, the restriction of \tilde{C}_R to $\mathcal{M}_I^{P \times K}$, $I \in \mathcal{I}$, is infinitely differentiable as a function of the nonzero entries of its matrix argument in $\mathcal{M}_I^{P \times K}$. Thus, one may look at the first and second derivatives of the restriction of \tilde{C}_R to $\mathcal{M}_I^{P \times K}$ to identify its local maximum points. The following result comes from Lemma 21 and the definition of \tilde{C}_R :

Lemma 22 For $I \in \mathcal{I}$, the restriction of $\tilde{\mathcal{C}}_R$ to $\mathcal{M}_I^{P \times K}$ admits the first and second partial derivatives

$$\begin{aligned} \frac{\partial \tilde{\mathcal{C}}_R(\mathbf{W})}{\partial W_{ij}} &= [\mathbf{W}^+]_{ji} - \frac{\text{sign}(W_{ij})R(S_j)}{\sum_{l=1}^K |W_{il}|R(S_l)}, \quad (i, j) \in I \\ \frac{\partial^2 \tilde{\mathcal{C}}_R(\mathbf{W})}{\partial W_{ij} \partial W_{kl}} &= [(\mathbf{WW}^T)^{-1}]_{ki}(\delta_{jl} - [\mathbf{W}^+ \mathbf{W}]_{lj}) - [\mathbf{W}^+]_{li}[\mathbf{W}^+]_{jk}, \\ &\quad (i, j) \in I, (k, l) \in I, k \neq i, \\ \frac{\partial^2 \tilde{\mathcal{C}}_R(\mathbf{W})}{\partial W_{ij} \partial W_{il}} &= \frac{\text{sign}(W_{ij})\text{sign}(W_{il})R(S_j)R(S_l)}{[\sum_{k=1}^K |W_{ik}|R(S_k)]^2} \\ &\quad + [(\mathbf{WW}^T)^{-1}]_{ii}(\delta_{jl} - [\mathbf{W}^+ \mathbf{W}]_{lj}) - [\mathbf{W}^+]_{li}[\mathbf{W}^+]_{ji}, \\ &\quad (i, j) \in I, (i, l) \in I. \end{aligned}$$

The above lemma allows one to characterize the stationary points of the restriction of $\tilde{\mathcal{C}}_R$ to $\mathcal{M}_I^{P \times K}$, by setting its derivative to zero, yielding

$$[\mathbf{W}^+]_{ji} = \frac{\text{sign}(W_{ij})R(S_j)}{\sum_{l=1}^K |W_{il}|R(S_l)}, \quad (i, j) \in I. \quad (3.65)$$

Thus one gets the following corollary.

Corollary 17 Let $I \in \mathcal{I}$, then for any $\mathbf{W} \in \mathcal{M}_I^{P \times K}$ which is a stationary point of the restriction of $\tilde{\mathcal{C}}$ on $\mathcal{M}_I^{P \times K}$:

$$\{(i, j) \in \mathbb{Z}^{P \times K} : [\mathbf{W}^+]_{ji} \neq 0\} \supseteq I,$$

and

$$\frac{\partial^2 \tilde{\mathcal{C}}_R(\mathbf{W})}{\partial W_{ij} \partial W_{il}} = [(\mathbf{WW}^T)^{-1}]_{ii}(\delta_{jl} - [\mathbf{W}^+ \mathbf{W}]_{lj}), \quad (i, j) \in I, (i, l) \in I.$$

The first statement of the corollary results directly from the fact that if $\mathbf{W} \in \mathcal{M}_I^{P \times K}$ then $W_{ij} \neq 0$ if $(i, j) \in I$ and both sides of Eq. (3.65) are non zero, too. The second claim is a consequence of the stationary point requirement: $\frac{\partial \tilde{\mathcal{C}}_R(\mathbf{W})}{\partial W_{ij}} = \frac{\partial \tilde{\mathcal{C}}_R(\mathbf{W})}{\partial W_{kl}} = 0$, $(i, j) \in I$, $(k, l) \in I$ (see Eq. (3.65)). Consider now the following lemma, proved in Section 3.8.14 (p. 185).

Lemma 23 Let $I \in \mathcal{I}$ such that either i) the set $\cup_{i=1}^P I_i$ contains more than P elements, or ii) there exists a pair of indices i, j in $\{1, \dots, P\}$ for which the $I_i \cap I_j \neq \emptyset$. Then the restriction of $\tilde{\mathcal{C}}_R$ in $\mathcal{M}_I^{P \times K}$ does not have a local maximum point.

Lemma 23 allows one to eliminate subsets I in \mathcal{I} for which the restriction of \mathcal{C}_R in $\mathcal{M}_I^{P \times K}$ does not have a local maximum point. It can be proved that the only subsets I of \mathcal{I} left are the ones such that all their i -th sections, $i = 1, \dots, P$,

are disjoint and reduce to a single point, which is the form of the non-mixing matrices forming the set $\mathcal{W}^{P \times K}$. But we know from Theorem 16 (p. 64) that these matrices necessarily induce a local maximum of $\tilde{\mathcal{C}}_R(\mathbf{W})$. This yields the discriminacy property of \mathcal{C}_R over $\mathcal{M}^{P \times K}$, which ensures the source recovering via the *local* maximization of $\mathcal{C}_R(\mathbf{B})$ (see the proof in Section 3.8.15, p. 186).

Corollary 18 (Discriminacy of $\mathcal{C}_R(\mathbf{B})$, partial approach) *The local maximum points of \mathcal{C}_R over the set $\mathcal{M}^{P \times K}$ correspond to \mathbf{B} such that $\mathbf{BA} \in \mathcal{W}^{P \times K}$ or, equivalently, $\mathbf{B} \sim_u \mathbf{A}^{-1}$.*

One concludes that for the specific $P = K$ case, $\mathcal{C}_R(\mathbf{B})$ admits a global maximum point \mathbf{B} if and only if $\mathbf{BA} \in \mathcal{W}^{K \times K}$ (note that if $P = K$, $\mathbf{W}^+ = \mathbf{W}^{-1}$), and one gets the main result of [Pham and Vrins, 2006]. The above results state that this result still holds for $P \leq K$: \mathcal{C}_R admits a local maximum point \mathbf{B} if (Theorem 16) and only if (Theorem 18) $\mathbf{BA} \in \mathcal{W}^{P \times K}$ and a global maximum point \mathbf{B} if and only if $\mathbf{BA} \in \mathcal{W}_P^{P \times K}$ (Corollary 6).

3.4.5 Jacobi updates, Givens trajectories and Discriminacy property

The discriminacy property of the range-based D-BSS, S-BSS and P-BSS contrast functions have been established in the previous subsections. These results still hold if one requires $\mathbf{B} \in \mathcal{SO}(K)$. More precisely, whatever is $\mathbf{B} \in \mathcal{SO}(K)$, the neighborhood of \mathbf{B} restricted to $\mathcal{SO}(K)$ is included in the neighborhood of \mathbf{B} in the whole space of square matrices. In other words, we can say that if a contrast admits a local maximum point at \mathbf{B} in the whole space, then the orthogonal counterpart of this contrast also has a local maximum at this point in $\mathcal{SO}(K)$. It was also stated that, conversely, the criterion is discriminant even if $\mathbf{B} \in \mathcal{SO}(K)$. However, this result supposes that in practice, the entire neighborhood of \mathbf{B} be explored in order to check if \mathbf{B} is indeed a local maximum point. Taking the S-BSS contrast as an example, this is clearly not the case if the orthogonal contrast is maximized using Jacobi updates (product of Givens matrices) of the form given in (1.67) where $\mathbf{R}^{(t)}$ is a Givens matrix. Indeed, for this update rule, only specific trajectories on the $\mathcal{SO}(K)$ manifold are permitted, they are not arbitrary on the associated manifold because the method proceeds by iterative rotations in the hyper-planes spanned by the pairs $(\mathbf{w}_p, \mathbf{w}_q)$ with $p < q$. In other words, some trajectories are unreachable. Only updates of the form $\mathbf{w}_p \leftarrow \cos \theta \mathbf{w}_p + \sin \theta \mathbf{w}_q$, $\mathbf{w}_q \leftarrow \cos \theta \mathbf{w}_q - \sin \theta \mathbf{w}_p$ are made possible. Therefore, along Jacobi trajectories, a contrast could seem to have a local maximum at a given point, even if this contrast can be increased along another trajectory of the restriction of $\mathcal{O}(K)$ to $\mathcal{SO}(K)$, as illustrated on the toy contrast shown in Figure 3.27. If, from a given point of $\mathcal{SO}(K)$, the contrast can be increased along a Jacobi trajectory, the matrix \mathbf{B} can be modified using the associated update. But otherwise, the fact that no Jacobi update leads to an increase of the contrast does not ensure, based on the above results, that we are in a local maximum point. Actually, the question that we would like to answer is the following: does there *always* exist a Givens trajectory such that the contrast can be increased if we are not

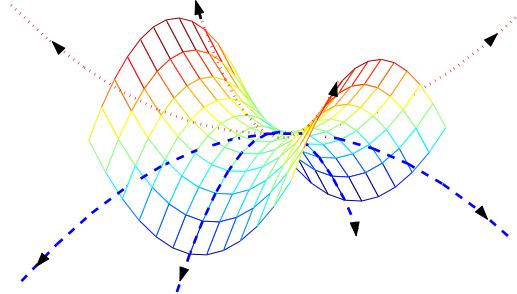


Figure 3.27. Toy contrast : an algorithm with a restricted field of geodesic trajectories may be stuck in a saddle point.

in a “true” local maximum in $\mathcal{SO}(K)$? We shall address this question by using a deflation procedure, where one is trying to extract a p -th source by updating the p -th row \mathbf{w}_p of \mathbf{W} .

As the first and second derivatives of $R(\mathbf{w}\mathbf{S}/\text{Var}[\mathbf{w}\mathbf{S}])$ are given in Lemma 17 (p. 149), one can compute the pushforward term of $f(\mathbf{w}_p) \doteq -R(\mathbf{w}_p\mathbf{S}/\text{Var}[\mathbf{w}_p\mathbf{S}])$ if $\mathbf{w}_p \leftarrow \mathbf{w}_p + \delta\mathbf{w}$, where $\delta\mathbf{w}$ is a small increment of \mathbf{w}_p . We have

$$f(\mathbf{w}_p + \delta\mathbf{w}) - f(\mathbf{w}_p) = \Delta + O(\|\delta\mathbf{w}\|^2) , \quad (3.66)$$

with $\Delta \doteq \langle \nabla_{\mathbf{w}_p} f(\mathbf{w}_p), \delta\mathbf{w} \rangle$. Remind that the first $p-1$ entries of \mathbf{w}_p are supposed to be zero and must remain so as the first $p-1$ outputs have already converged to the first $p-1$ sources. Hence, any suitable Jacobi update can be written as

$$\mathbf{w}_p \leftarrow \mathbf{w}_p + \underbrace{\mathbf{w}_p(\cos \theta - 1) + \mathbf{w}_q \sin \theta}_{\doteq \delta\mathbf{w}} , \quad (3.67)$$

with $q > p$ and $\delta\mathbf{w} = \delta\mathbf{w}(\theta)$. Because $I(\mathbf{w}_p) = I(\mathbf{w}_q)$ does not necessarily hold true, we may face a problem. To compute the pushforward Δ of f in this direction $\delta\mathbf{w}(\theta)$, we need the gradient of $f(\mathbf{w}_p)$ in this direction, which does not exist for the index entries $i \in I(\mathbf{w}_q) \setminus I(\mathbf{w}_p)$. However, this gradient exists “right near” the zero element: the k -th entry of the gradient is (if $\mathbf{w}_p(k) = 0$) $\lim_{\mathbf{w}_p(k) \uparrow 0} \nabla_{\mathbf{w}_p} f(\mathbf{w}_p) = R(\mathbf{S}_k)/\|\mathbf{w}_p\|$ or $\lim_{\mathbf{w}_p(k) \downarrow 0} \nabla_{\mathbf{w}_p} f(\mathbf{w}_p) = -R(\mathbf{S}_k)/\|\mathbf{w}_p\|$, depending on the sign of the k -th entry of the increment $\delta\mathbf{w}(\theta)$. In other words, the pushforward term Δ is the sum of two terms: the first one (T_1) results from a change of direction along the non-zero-elements of \mathbf{w}_p , and the second one (T_2) from a change of subspace (intuitively, we “leave” the subspace corresponding to the zero-elements of \mathbf{w}_p towards a higher dimensional subspace). Defining

$\bar{I}(\mathbf{w}_p) \doteq I(\mathbf{w}_q) \setminus I(\mathbf{w}_p)$ for short, we have (noting that $\|\mathbf{w}_p\| = 1$) :

$$\begin{aligned} T_1 &= - \sum_{i \in I(\mathbf{w}_p)} \{ (\cos \theta - 1) (|\mathbf{w}_p(i)|R(\mathbf{S}_i) - \mathbf{w}_p(i)^2 R(\mathbf{w}_p \mathbf{S})) \\ &\quad + \sin \theta \mathbf{w}_q(i) (\text{sign}(\mathbf{w}_p(i))R(\mathbf{S}_i) - \mathbf{w}_p(i)R(\mathbf{w}_p \mathbf{S})) \} \\ T_2 &= -\sin \theta \sum_{i \in \bar{I}(\mathbf{w}_p)} \mathbf{w}_q(i) \text{sign}(\mathbf{w}_q(i) \sin \theta) R(\mathbf{S}_i) . \end{aligned}$$

Then, $\Delta = T_1 + T_2$ reduces to:

$$\begin{aligned} \Delta &= -\theta \left\{ \underbrace{\sum_{i \in I(\mathbf{w}_p)} \mathbf{w}_q(i) [\text{sign}(\mathbf{w}_p(i))R(\mathbf{S}_i) - \mathbf{w}_p(i)R(\mathbf{w}_p \mathbf{S})]}_{\doteq T'_1(q)} \right. \\ &\quad \left. + \underbrace{\sum_{i \in \bar{I}(\mathbf{w}_p)} \text{sign}(\mathbf{w}_q(i)\theta) \mathbf{w}_q(i) R(\mathbf{S}_i)}_{\doteq T'_2(q,\theta)} \right\} + O(\theta^2) \quad (3.68) \\ &\doteq \Delta_\theta + O(\theta^2) \end{aligned}$$

Observe that $\Delta_\theta + O(\theta^2)$ also corresponds to $\Delta + O(\|\delta \mathbf{w}(\theta)\|^2)$, since $\|\delta \mathbf{w}(\theta)\|^2 = \theta^2 + O(\theta^4)$ is $O(\theta^2)$ (as $\theta \rightarrow 0$). This means intuitively that $\|\delta \mathbf{w}(\theta)\|^2$ tends to zero at least as fast as a constant times θ^2 . In other words, the higher-order terms in θ can be relegated in the $O(\|\delta \mathbf{w}(\theta)\|^2)$ term, so that one can replace Δ by Δ_θ in Eq. (3.68) without affecting the equality. Mathematically, this is because $\Delta + O(\|\delta \mathbf{w}(\theta)\|^2) = \Delta_\theta + O(\theta^2) + O(\|\delta \mathbf{w}(\theta)\|^2) = \Delta_\theta + O(\|\delta \mathbf{w}(\theta)\|^2)$.

Let us analyze this pushforward term Δ_θ .

1. First, note that if $\# [I(\mathbf{w}_p)] = 1$, then $\mathbf{w}_q(i) = 0$ for all $i \in I(\mathbf{w}_p)$ and all $p < q \leq K$ because $\mathbf{w}_p \mathbf{w}_q^\top = 0$; consequently, $T'_1(q) = 0$ for all $p < q \leq K$. But since $\text{sign}(ab) = \text{sign}(a)\text{sign}(b)$ and $a\text{sign}(a) = |a|$, we have $\Delta_\theta < 0$; we are in a local maximum of f , as expected.
2. Assume now that $\bar{I}(\mathbf{w}_p) = \emptyset$. Note that

$$T'_2(q, \theta) = \text{sign}(\theta) \sum_{i \in \bar{I}(\mathbf{w}_p)} |\mathbf{w}_q(i)| R(\mathbf{S}_i) .$$

Consequently, for $\theta \neq 0$, $T'_2(q, \theta) = 0$ is equivalent to $\bar{I}(\mathbf{w}_p) = \emptyset$ (intuitively: we stay within the same subspace). Then $\Delta_\theta \propto \theta T'_1(q)$. Except if $T'_1(q) = 0$ as well, the pushforward can be made positive by choosing the right sign for θ . What if $T'_1(q) = 0$? This may result from two cases:

- *the gradient vanishes.* This may happen at the local minimum of $f(\mathbf{w}_p)$: the stationary point condition of f (given in Eq. (3.57)) is the same as imposing that each term into the brackets in $T'_1(q)$ (which are exactly the entries of $\nabla_{\mathbf{w}_p} f(\mathbf{w}_p)$) is zero. There is thus no problem.
 - *the restricted gradient (short-hand reference for “the subvector of $\nabla_{\mathbf{w}_p} f(\mathbf{w}_p)$ with respect to $I(\mathbf{w}_p)$ ” does not vanish but is orthogonal to any of the last subvectors $\widehat{\mathbf{w}}_{p+1}, \dots, \widehat{\mathbf{w}}_K$ of $\mathbf{w}_{p+1}, \dots, \mathbf{w}_K$ (respectively) with respect to $I(\mathbf{w}_p)$ (see Def. 24 for a definition of “subvectors”).* This cannot happen but, as it could be feared, the justification is more involved. By definition of $I(\mathbf{w}_p)$, this case is not possible because simultaneously i) $\widehat{\mathbf{w}}_p$ is perpendicular to these subvectors, since the null entries of \mathbf{w}_p do not influence the value of the dot product: $\mathbf{w}_p \mathbf{w}_q^T = \widehat{\mathbf{w}}_p \widehat{\mathbf{w}}_q^T$, ii) the restricted gradient is not co-linear with $\widehat{\mathbf{w}}_p$ (it is easily seen that it is perpendicular) and iii) the set $\{\widehat{\mathbf{w}}_p, \dots, \widehat{\mathbf{w}}_K\}$ spans a $\#[I(\mathbf{w}_p)]$ -dimensional space; it also spans all the subspaces of dimensions lower than $K - p + 1$, and in particular, the one corresponding to the dimension of the gradient of $\widehat{\mathbf{w}}_p$, that is of dimension $\#[I(\mathbf{w}_p)] \leq K - p + 1$. In other words, iii) results from the fact that any $K - p + 1$ -entry vector (and in particular, the restricted gradient of f with respect to $I(\mathbf{w}_p)$) can be written as a linear combination $\sum_{i=1}^{K-p+1} \alpha(i) \widehat{\mathbf{w}}_{p+i-1}$. As the coefficients $\alpha(i)$ cannot all vanish for non-null vectors (and the gradient is assumed to not vanish), the dot product between the restricted gradient and each of the last $K - p$ subvectors $\widehat{\mathbf{w}}_{p+1}, \dots, \widehat{\mathbf{w}}_K$ cannot vanish.
- ⇒ The $\Delta_\theta = 0$ corresponding to $T'_2(q, \theta) = T'_1(q) = 0$ for all $q > p$ and all small enough angles θ results from the fact that the gradient $\nabla_{\mathbf{w}_p} f(\mathbf{w}_p)$ vanishes, and that we are in a local minimum of the function: one can always increase the function by applying a sufficiently small angular variation. On the other hand, if for at least one q satisfying $\bar{I}(\mathbf{w}_p) = \emptyset$ we have $T'_1(q) \neq 0$, then the pushforward can be made positive, too.

3. Assume now $\bar{I}(\mathbf{w}_p) \neq \emptyset$. Do we necessarily have that $\Delta_\theta > 0$ or $\Delta_{-\theta} > 0$ for at least one row-index $q > p$? This is a necessary condition to ensure that the criterion does not reach a mixing stationary point or mixing local maximum point. From Eq. (3.68) we have

$$\text{sign}(\Delta_\theta) = -\text{sign}(\theta) \text{sign}(T'_1(q) + T'_2(q, \theta)) . \quad (3.69)$$

Using the equality $\widehat{\mathbf{w}}_q \widehat{\mathbf{w}}_p^T = 0$, this necessary condition is equivalent to show that there exists $\delta > 0$ and at least one index $q \in \{p+1, \dots, K\}$ such

that $\text{sign}(\theta)$ equals

$$-\text{sign}\left(\sum_{i \in I(\mathbf{w}_p)} \mathbf{w}_q(i) \text{sign}(\mathbf{w}_p(i)) R(\mathbf{S}_i) + \text{sign}(\theta) \sum_{i \in \bar{I}(\mathbf{w}_p)} |\mathbf{w}_q(i)| R(\mathbf{S}_i)\right) \quad (3.70)$$

for all $|\theta| < \delta$. This is not necessarily the case, as shown in the following example. Note that it is not a “random” counter-example. It has been chosen according to the above observation saying that some problem might be encountered if $\bar{I}(\mathbf{w}_p) \neq \emptyset$ for all $q > p$.

Example 25 Since the first extracted rows have no importance, we can set without loss of generality $p = 1$. Assume $R(\mathbf{S}_i) = 1$ for all i . The following 3×3 orthogonal transfer matrix yields a counter-example showing that the push-forward term may be negative along all Jacobi trajectories:

$$\mathbf{W} = \begin{bmatrix} 0 & -0.8 & -0.6 \\ 0.8 & -0.36 & 0.48 \\ -0.6 & -0.48 & 0.64 \end{bmatrix}. \quad (3.71)$$

The only possible rotations leading to the update rule (3.67) are obtained by considering the first row of \mathbf{W} after left-multiplication by a Givens matrix $\mathbf{G}_{12}^{\pm\theta}$ ($q = 2$) or $\mathbf{G}_{13}^{\pm\theta}$ ($q = 3$). Let us denote the corresponding new transfer matrices by \mathbf{W}' and \mathbf{W}'' , respectively. Figure 3.28.(a) shows $f(\mathbf{w}_1)$ (dotted line), $f(\mathbf{w}'_1)$ (solid line) and $f(\mathbf{w}''_1)$ (dashed line) as a function of θ ; the simulation was performed for 20 points θ equally spaced in $[-0.5, 0.5]$. It is seen that the criterion decreases in both cases whatever the sign of θ while we are not in a true local maximum if the whole search space is considered, as explained above. The sceptic reader could doubt about the result shown in this figure, because we have not evaluated the function f “right near $\theta = 0$ ”, but for some $|\theta| > 0$: theoretically speaking, there could exist θ' satisfying $0 < |\theta'| < |\theta|$ such that the function is increased for the angular increment θ' . In this sense, we agree that Fig. 3.28.(a) does not constitute an absolute proof (no limit for $|\theta| \rightarrow 0$ has been computed). However, a rigorous proof is given by Fig. 3.28.(b) in which the term $\text{sign}(T'_1(q) + T'_2(q, \theta))$ in the right-hand side of Eq. (3.69) (which depends on the sign of θ only, not on the value of θ) has been plotted vs θ : for both $q = 2$ and $q = 3$, these terms equal $\text{sign}(\theta)$, i.e. by Eq. (3.69), this leads to $\Delta_\theta < 0$ for all $q > p$ and all “sufficiently small” rotation angles. Note that if “simultaneous rotations” are possible (and not only successive rotations in various hyper-planes), i.e. when rotation matrices other than Givens matrices can be used to perform the geodesic optimization over $\mathcal{SO}(K)$, this does not happen. For example, if \mathbf{W} is left multiplied by $\mathbf{R} \doteq \mathbf{G}_{12}^\theta \mathbf{G}_{13}^\theta$, the criterion increases for sufficiently small $\theta > 0$ while we still have $\mathbf{RW} \in \mathcal{SO}(K)$.

In conclusion, the above reasoning shows that spurious maxima of the range-based contrast function may be faced along Givens trajectories. The above analysis shows that this is possible when we have to “leave” the subspace spanned by the basis vectors corresponding to the non-zero entries.

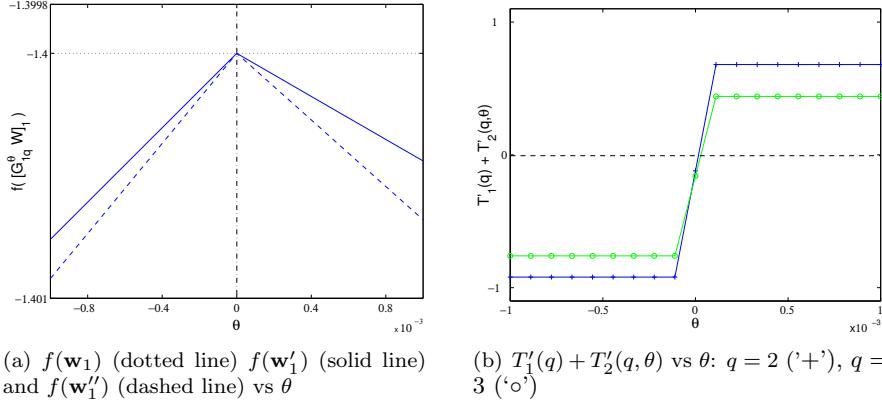


Figure 3.28. Example 25. The range-based contrast function is not discriminant along Jacobi trajectories.

3.5 DISCRIMINACY OF THE MINIMUM SUPPORT (ZERO-ENTROPY) APPROACH

It has been explained in Chapter 2 that both the support and the range can be used as cost functions for the separation of bounded sources. There is no reason, a priori, to prefer the support to the range functional. On the contrary, estimating the support should be more complicated than estimating the range as the estimation of the latter is necessary to estimate the former. Another problem regarding the support compared to the range is the problem of mixing maxima. It has been proved in the previous section that the range functional has no mixing maxima, whatever the extraction scheme (deflation, simultaneous or partial techniques); this is an additional advantage of the range compared to the support since, as it will be shown below, this property is not shared by the support. Because of the number of disadvantages of the support compared to the range, we restrict ourselves to give a simple $K = 2$ example of sources sharing a same not necessarily symmetric density p_S for which mixing maxima can be observed. Indeed, discussing all the conditions on the source supports ensuring the existence of such maxima is of few interest in practice since the range should always be preferred to the support criterion.

Inspired by Lemma 6 (p. 60) which states a difference of behavior between the support and the range, we choose for p_S a suitable density with non-convex support. Let p_{S_1} and p_{S_2} be two densities of independent random variables $S_i = U_i + D_i$ where U_1 and U_2 are independent uniform variables taking non-zero values in $[-\nu, \nu]$ ($\nu > 0$) and D_1, D_2 are independent discrete random variables taking values $[\alpha, 1 - \alpha]$ at $\{-\xi, \xi\}$ ($\xi > 0$). Suppose further that $\xi > 2\nu$. Then,

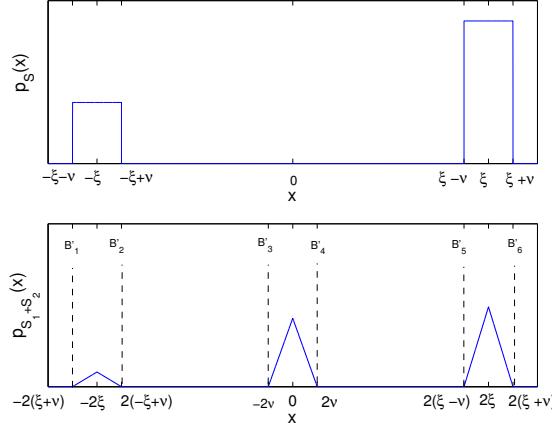


Figure 3.29. Source density p_S and pdf of $p_{S_1+S_2}$.

both sources S_i have the same density p_S (see Fig. 3.29.):

$$p_S(s) = \begin{cases} \frac{\alpha}{2\nu} & \text{for } s \in [-\xi - v, -\xi + v] \\ \frac{1-\alpha}{2\nu} & \text{for } s \in [\xi - v, \xi + v] \\ 0 & \text{elsewhere.} \end{cases} \quad (3.72)$$

It results that $\Omega(S_i) = \{x \in [-\xi - v, -\xi + v] \cup [\xi - v, \xi + v]\}$ and $\bar{\Omega}(S_i) = \{x \in [-\xi - v, \xi + v]\}$, which implies $\mu[\Omega(S_i)] = 4v$ and $\mu[\bar{\Omega}(S_i)] = R(S_i) = 2\xi + 2v$. By Lemma 6, we have $\mu[\bar{\Omega}(S_1 + S_2)] = \mu[\bar{\Omega}(S_1)] + \mu[\bar{\Omega}(S_2)] = 8v$ and $\mu[\Omega(S_1 + S_2)] > \mu[\Omega(S_1)] + \mu[\Omega(S_2)]$ ($\mu[\Omega(S_1 + S_2)] = 12v$). This can be observed in Figure 3.29. The interval bounds $B'_1 - B'_6$ can be easily found by an intuitive reasoning but also by pointing out that the density $p_{S_1+S_2}$ is obtained by the convolution of p_{S_1} and p_{S_2} by the independence assumption on the sources. As $\xi > 2v$ (i.e. $\xi - v > v$), the support of $\Omega(S_1 + S_2)$ is composed of 3 disjoint intervals $(B'_1, B'_2), (B'_3, B'_4), (B'_5, B'_6)$.

We shall prove that if $\|\mathbf{w}\|$ is kept fixed (here: equal to one), the criterion $\mu[\Omega(\mathbf{w}S)]$ reaches a local maximum point at $\mathbf{w} = \mathbf{w}_{\pi/4}$, which does not yield $\mathbf{w}_{\pi/4}S \propto S_i$. The question that has to be dealt with is the following: does there exist $\delta > 0$ such that $\forall \Delta > 0$ s.t. $|\Delta| < \delta$, $\mu[\Omega(\mathbf{w}_{\pi/4+\Delta}S)] > \mu[\Omega(\mathbf{w}_{\pi/4}S)]$?

According to the Lebesgue measure definition, the measure of union of *disjoint* intervals is the sum of the interval lengths. Therefore, $\mu[\Omega(\mathbf{w}_{\pi/4}S)]$ equals $(B''_2 - B''_1) + (B''_4 - B''_3) + (B''_6 - B''_5)$ where the B''_i are given by $\sqrt{2}/2B'_i$ (this is because $\mathbf{w}_{\pi/4+\Delta}S$ is a scaled version of $p_{S_1+S_2}$: the x axis scale is expanded by a factor $\sqrt{2}/2$ and the y axis by its inverse and therefore, the support of $\mathbf{w}_{\pi/4+\Delta}S$ is also composed of three disjoint intervals, with bounds B''_i given right above,

see Fig. 3.29.) Remind that $\mu[\Omega(\mathbf{w}_{\pi/4}\mathcal{S})] = (\sqrt{2}/2)12\nu = 6\sqrt{2}\nu$ and note that $\mathbf{w}_{\pi/4+\Delta} = \sqrt{2}/2[\sin \Delta + \cos \Delta, \cos \Delta - \sin \Delta]$.

Because $\xi - \nu > \nu$, for sufficiently small Δ , the support of $\mathbf{w}_{\pi/4+\Delta}\mathcal{S}$ is still composed of three disjoint intervals. In order to compute $\mu[\Omega(\mathbf{w}_{\pi/4+\Delta}\mathcal{S})]$, we shall look at the bounds B_1, B_2, B_3, B_4, B_5 and B_6 of the above three intervals, that are defined similarly as in Fig. 3.29. The sought quantity would reduce to $(B_2 - B_1) + (B_4 - B_3) + (B_6 - B_5)$.

Whatever is θ , the support of $\mathbf{w}_\theta\mathcal{S}$ can be written as the union of four (not necessarily disjoint) intervals Ω_i , $1 \leq i \leq 4$. They are built from the pairwise convolution of the pair of rectangles in the pdfs $p_{\mathbf{w}_\theta(1)\mathcal{S}_1}$ and $p_{\mathbf{w}_\theta(2)\mathcal{S}_2}$ (because both $p_{\mathbf{w}_{\pi/4+\Delta}(1)\mathcal{S}_1}$ and $p_{\mathbf{w}_{\pi/4+\Delta}(2)\mathcal{S}_2}$ are a scaled version of $p_{\mathcal{S}}$, they also have the same shape as $p_{\mathcal{S}}$ for $\theta \notin \{k\pi/2k \in \mathbb{Z}\}$): left-left, left-right, right-left, right-right yielding the intervals $\Omega_1, \Omega_2, \Omega_3$ and Ω_4 , respectively.

The interval $\Omega_1 = (B_1, B_2)$ results from the left-left combination and $\Omega_4 = (B_5, B_6)$ from the right-right one. It is easy to see that $\sin(\pi/4 + \Delta) + \cos(\pi/4 + \Delta) = \sqrt{2} \cos \Delta$ and, from Fig. 3.29. :

- $B_1 = (\mathbf{w}_{\pi/4+\Delta}(1) + \mathbf{w}_{\pi/4+\Delta}(2))(-\xi - \nu) = \sqrt{2} \cos \Delta(-\xi - \nu)$;
- $B_2 = (\mathbf{w}_{\pi/4+\Delta}(1) + \mathbf{w}_{\pi/4+\Delta}(2))(-\xi + \nu) = \sqrt{2} \cos \Delta(-\xi + \nu)$;
- $B_5 = (\mathbf{w}_{\pi/4+\Delta}(1) + \mathbf{w}_{\pi/4+\Delta}(2))(\xi - \nu) = \sqrt{2} \cos \Delta(\xi - \nu)$;
- $B_6 = (\mathbf{w}_{\pi/4+\Delta}(1) + \mathbf{w}_{\pi/4+\Delta}(2))(\xi + \nu) = \sqrt{2} \cos \Delta(\xi + \nu)$.

This gives $B_2 - B_1 = B_6 - B_5 = 2\nu\sqrt{2} \cos \Delta$.

Let us now focus on the remaining intervals Ω_2 and Ω_3 . If $\Delta = 0$, then $\Omega_2 = \Omega_3$ (as in the bottom of Fig. 3.29.), but for a small $\Delta \neq 0$, the above two intervals are different, though not disjoint. Then, the computation of $(B_3, B_4) = \Omega_2 \cup \Omega_3$ requires more attention, since it is the union of 2 overlapping intervals Ω_i . Actually, it is easily checked that

$$\begin{aligned} B_3 &= \min \left(\mathbf{w}_{\pi/4+\Delta}(1)(-\xi - \nu) + \mathbf{w}_{\pi/4+\Delta}(2)(\xi - \nu), \right. \\ &\quad \left. \mathbf{w}_{\pi/4+\Delta}(1)(\xi - \nu) + \mathbf{w}_{\pi/4+\Delta}(2)(-\xi - \nu) \right) \\ &= \sqrt{2} \min (-\xi \sin \Delta - \nu \cos \Delta, \xi \sin \Delta - \nu \cos \Delta) \\ &= \sqrt{2}(-\xi |\sin \Delta| - \nu \cos \Delta) \end{aligned} \tag{3.73}$$

and

$$\begin{aligned} B_4 &= \max \left(\mathbf{w}_{\pi/4+\Delta}(1)(\nu - \xi) + \mathbf{w}_{\pi/4+\Delta}(2)(\xi + \nu), \right. \\ &\quad \left. \mathbf{w}_{\pi/4+\Delta}(1)(\xi + \nu) + \mathbf{w}_{\pi/4+\Delta}(2)(\nu - \xi) \right) \\ &= \sqrt{2} \max (-\xi \sin \Delta + \nu \cos \Delta, \xi \sin \Delta + \nu \cos \Delta) \\ &= \sqrt{2}(\xi |\sin \Delta| + \nu \cos \Delta) \end{aligned} \tag{3.74}$$

i.e. $B_4 - B_3 = 2\sqrt{2}(\xi|\sin \Delta| + \nu \cos \Delta)$.

As for sufficiently small Δ we have $B_2 < B_3$ and $B_4 < B_5$, the support is composed of three disjoint interval. Therefore, $\mu[\Omega(\mathbf{w}_{\pi/4+\Delta}\mathbf{S})]$ equals the sum $(B_2 - B_1) + (B_4 - B_3) + (B_6 - B_5)$:

$$\mu[\Omega(\mathbf{w}_{\pi/4+\Delta}\mathbf{S})] = 6\sqrt{2}\nu \cos \Delta + 2\sqrt{2}\xi|\sin \Delta| , \quad (3.75)$$

and it can be checked that $\lim_{\Delta \rightarrow 0} \mu[\Omega(\mathbf{w}_{\pi/4+\Delta}\mathbf{S})] = 6\sqrt{2}\nu = \mu[\Omega(\mathbf{w}_{\pi/4}\mathbf{S})]$ as it should. Considering a first order expansion of the above circular functions, we finally get that for sufficiently small $|\Delta| > 0$:

$$\begin{aligned} \mu[\Omega(\mathbf{w}_{\pi/4+\Delta}\mathbf{S})] &= 6\nu\sqrt{2} + 2\sqrt{2}\xi|\Delta| + o(\Delta) \\ &\geq \mu[\Omega(\mathbf{w}_{\pi/4}\mathbf{S})] \end{aligned} \quad (3.76)$$

with equality if and only if $\Delta = 0$. This shows that the criterion (resp. $-\mu[\Omega(\mathbf{w}_\theta\mathbf{S})]$) has a local minimum (resp. maximum) point at $\theta = \pi/4$.

Remark 21 *The simultaneous support-based orthogonal contrast, still with $K = 2$, reduces to*

$$-\mu[\Omega(\mathbf{w}_\theta\mathbf{S})] - \mu[\Omega(\mathbf{w}_{\pi/2-\theta}\mathbf{S})] . \quad (3.77)$$

Clearly, in the above results, the parameter α used in Eq. (3.72) has no matter as long as it belongs to $(0, 1)$. Then, we can set without loss of generality $\alpha = 0.5$, implying that the common source density is symmetric. This gives that the support of the first output $\sin \theta S_1 + \cos \theta S_2$ equals the support of the second one $-\cos \theta S_1 + \sin \theta S_2$ as they share a same symmetric pdf (the convolution of two symmetric functions is symmetric and ‘‘reversing’’ one of these function does not affect the convolution product). Therefore, the above criterion reduces to $-2\mu[\Omega(\mathbf{w}_\theta\mathbf{S})]$, which proves that the simultaneous orthogonal criterion also has a local maximum point at $\theta = \pi/4$.

Part of the above discussion first appeared in [Vrins et al., 2006].

3.6 SUMMARY OF RESULTS AND CONTRAST SETS CONFIGURATION

Based on the results of Chapter 2 and Chapter 3, we can find the right configuration in Figure 3.1.

- In Chapter 2, it was proved that $\mathcal{C}_h(\cdot) \in \tilde{\mathbb{F}}$. We have seen in Section 3.2 that mixing local maximum points may exist for some source densities; we conclude that $\mathcal{C}_h(\cdot) \notin \mathbb{F}_{\mathcal{C}^D}$. This proves that $\mathbb{F}_{\mathcal{C}^D} \not\subseteq \tilde{\mathbb{F}}$ and consequently, neither Conf.1 nor Conf.2 are correct.
- Contrary to Shannon’s entropy-based contrasts, $\mathcal{C}_R(\cdot) \in \tilde{\mathbb{F}}$ and $\mathcal{C}_R(\cdot) \in \mathbb{F}_{\mathcal{C}^D}$, which shows that $\mathbb{F}_{\mathcal{C}^D} \cap \tilde{\mathbb{F}} \neq \emptyset$, and Conf. 5 has to be rejected.

Then, the acceptable configuration is either Conf. 3 or Conf. 4. Some orthogonal contrast functions, like the (sum of the) output absolute (or square) kurtosis are not in the set $\bar{\mathcal{F}}_C$; this results from the fact that the cumulants satisfy $\text{cum}_r(\alpha X) = |\alpha|^r \text{cum}_r(X)$ (the sensitivity to scaling is avoided thanks to the whitening preprocessing) and are strictly additive for any pair of independent random variables

$$\text{cum}_r(X + Y) = \text{cum}_r(X) + \text{cum}_r(Y) \quad (3.78)$$

and thus the r -th root of the absolute r -th cumulant is a class II r -subadditive functional:

$$\sqrt[r]{|\text{cum}_r(X + Y)|} \leq \sqrt[r]{|\text{cum}_r(X)|} + \sqrt[r]{|\text{cum}_r(Y)|}. \quad (3.79)$$

It is explained in [Comon, 1994] that the sum of the square cumulants of order r is a contrast function for $r \geq 3$ if at most one source has a zero r -th order cumulant.

In can be shown that they might be discriminant (see e.g. [Delfosse and Loubaton, 1995] for a proof regarding the deflation approach using the output square kurtosis or the cumulant-based sinusoidal contrast in [Murillo-Fuentes and Gonzalez-Serrano, 2004], and algebraic techniques exist for the maximization of their counterpart for simultaneous separation); hence, they are in \mathcal{F}_{CD} . This yields that the only acceptable configuration is Conf. 4.

We concludes the chapter with the following lemma, which is an extnsion of Pham's theorem (Theorem 8, p. 50) based on the range properties.

Lemma 24 *Let $Q(\cdot)$ be a positive real-valued functional defined on the space of one-dimensional random variables. Let $\Psi(\cdot)$ be a strictly increasing mapping. Assume that $Q(\cdot)$ is scale-equivariant and satisfies $Q^2(X + Y) \geq Q^2(X) + Q^2(Y)$. Then*

- $\Psi\left(\frac{\sqrt{\text{Var}[bX]}}{Q(bX)}\right)$, $\Psi\left(\frac{|\det \mathbf{B}|}{\prod_{i=1}^K Q(b_i X)}\right)$ and $\Psi\left(\frac{\det(\mathbf{B}\Sigma_X \mathbf{B}^T)}{\prod_{i=1}^P Q(b_i X)}\right)$ are deflation, simultaneous and partial contrast functions, respectively;
- if in addition $Q(X + Y) = Q(X) + Q(Y)$ holds true, then the above contrast functions are discriminant, in the sens of the terminology used in this thesis.

3.7 CONCLUSION OF THE CHAPTER

3.7.1 Summary of results

In this chapter, we have investigated the question of the possible existence of spurious maxima in entropy-based contrast functions: we have looked if some

local maximum point of entropic criteria can be reached for transfer vectors $\mathbf{w} \in \mathbb{R}^K$ not proportional to the basis vectors. It was first explained using an informal approach that Shannon's entropy-based (and also mutual information-based) criteria can suffer from this problem in the specific situation where the source pdfs are multimodal; the locations of these maxima in \mathbb{R}^K can be related to the vectors \mathbf{w} that correspond to a local minimum number of the modes of the output pdf. This has been rigorously proved using two different approaches: a Taylor expansion of the entropy (a necessary and sufficient condition for the existence of spurious maxima was provided in the specific $K = 2$ case when the two sources share a same symmetric density) and using entropy estimation with bounds. The last approach yields more general results, and the entropy approximator can be used in other contexts than BSS. In both cases, the relationship with the mode standard deviation of the multimodal source pdf was emphasized. Also, the mutual information suffers from the same drawback. It was noted in [Achard, 2003] that the gradient of the output mutual information may vanish even when the outputs are mutually dependent, but the characterization of this stationary point (local minima, local maxima, saddle point?) was lacking. In this chapter, it is rigorously proved that these points may correspond to local minimum points.

The discriminacy analysis of the general Rényi's entropy has not been investigated as, from our theoretical and experimental results, only Shannon's entropy and (extended) Rényi's entropy with $r = 0$ are, whatever the source densities, truly cost functions for BSS under a (feasible) fixed variance constraint; this was shown in Chapter 2.

Rényi's entropy with $r = 0$ has then been studied; a simple extension of Rényi's entropy has also been proposed. The last extension exactly matches the standard Rényi entropy except, possibly, regarding the 0-entropy. In spite of its simplicity, this trivial extension may, however, lead to a functional having a more desirable behavior than the usual Rényi entropy in this case, though. It is shown based on various approaches that the range criterion (extended Rényi's entropy with $r = 0$) yields a discriminant contrast function provided that the sources are bounded, whatever the extraction scheme: deflation, symmetric or partial separation. To our knowledge, the range-based criterion is the only existing contrast, up to now, that is proved to be discriminant (even without prewhitening step !) in these three scenarios. On the other hand, a simple example showed that the support-based contrast may have spurious maxima, even though it was shown in Chapter 2 that it is a contrast function: it suffers from the same drawback as Shannon's entropy.

The last lemma of the chapter also states a sufficient condition for a contrast to be discriminant. It suffices that it fulfills a specific form and a strict additivity condition; this is an extension of Pham's theorem (Theorem 8 p. 50). However the above section emphasizes the connection between the known discriminant contrast functions (kurtosis-based and range-based ones). They are mappings (e.g. absolute value) of strictly additive quantities (cumulants and range itself, respectively).

	Deflation	Simultaneous	Partial	Discriminacy
Shannon ($r = 1$)	OK	OK	OK	KO ⁺
Hartley ($r = 0$)	OK	OK	OK	KO
Rényi ($r > 0, r \neq 1$)	KO ⁻	KO ⁻	KO ⁻	NA
Range (ext. Hartley)	OK	OK	OK	OK

Table 3.2. Summary of the results of Chapter 2: analysis of the contrast property of entropy-based criteria for the deflation, simultaneous and partial BSS. It is rigorously proved that Shannon, Hartley and extended Hartley entropies all yield to contrast function for the three separation schemes. By contrast, it always exist counter-examples showing that Rényi's entropy might be not a contrast function whatever is $r > 0, r \neq 1$. Original results are boldfaced, alternative proofs have been used to prove the known results.

The results of this chapter are summarized in the last column of Table 3.2. The results proved in the other columns were proved in Chapter 2. The “KO” results are proved via theoretical counterexamples showing that the corresponding property might be violated in some cases (even under the usual non-Gaussianity assumption). The “⁻” superscript indicates that these results are unexpected (but not contradictory) compared to the literature, and the “⁺” superscript indicates that the results confirm previous numerical experiments involving approximations.

3.7.2 Comparison with existing results

Our result might be considered as in contradiction with the results of Babaie-Zadeh. Indeed, he claims in [Babaie-Zadeh, 2002] that the mutual information cannot have a local minimum if it is non-zero. Then, how could we face local minima of the mutual information when the outputs have not been separated, remain thus necessarily dependent and consequently, when the output mutual information is non-zero ? The apparent contradiction comes from the fact that in [Babaie-Zadeh, 2002], the author does not assume any model. To see that, let us give the original theorem:

Theorem 20 (Babaie-Zadeh) *Let \mathbf{X} be a random vector with a continuously differentiable pdf. If for any “small vector” Δ the inequality $KL(\mathbf{X}) \leq KL(\mathbf{X} + \Delta)$, then $KL(\mathbf{X}) = 0$.*

In other words, the above theorem says that if the outputs are not mutually independent, it is always possible to add a random vector to \mathbf{X} such that the output mutual information will decrease.

Note that there is no restriction on the random vector Δ . In particular, it is not constrained to be of the form $\alpha\mathbf{X}$ (the fact that there is no fixed variance constraint or other tricks preventing the outputs to converge to null signals is

not of primary importance as mutual information is not scale sensitive). On the contrary, our analysis focuses on specific random vectors, that can be written as linear combinations of the mixtures. This is because we are trying to recover independent source signals from linear mixtures of them; not only to produce *any* random vectors with independent components. Hence, the results are not contradictory. This was pointed out by Babaie-Zadeh and Jutten in [Babaie-Zadeh and Jutten, 2005].

An interesting question that would have to be explored is the condition (depending on both the contrast function and the sources) ensuring that no local optima exist. This is the converse problem of the one addressed in this chapter, since we have given sufficient conditions for the existence of such optima, focusing on multimodal source densities. A partial converse result exists: in 2003, Boscolo & Roychowdhury showed that in a specific $K = 2$ case, the mutual information local minimum is also global: the local minimum is unique (up to the usual indeterminacies) [Boscolo and Roychowdhury, 2003]. The remaining open problem is to reduce the gap (between the results in [Boscolo and Roychowdhury, 2003] and those provided in this chapter) in which “nothing can be said”, so far. On the one hand, we know from [Boscolo and Roychowdhury, 2003] that for two i.i.d., symmetric and “nearly Gaussian sources” (in the sense of the Gram-Charlier expansion of the source pdfs), the mutual information contrast is discriminant and on the other hand, we now that we shall face the existence of mixing optima of the mutual information when the sources are strongly multimodal... But what about the possibility to have a spurious optimum when considering other kinds of unimodal or smoothly multimodal densities? None of the approaches can currently answer this question.

3.8 APPENDIX: PROOFS OF RESULTS OF THE CHAPTER

3.8.1 Proof of relation (3.14) (wording p. 114)

The left-hand side of Eq. (3.14) can be written as $\log |\det((\mathbf{I}_K + \mathcal{E})\mathbf{B})|$, where it is assumed $\|\mathcal{E}\| \ll 1$. Let us denote the EVD decomposition of \mathcal{E} , $\text{EVD}(\mathcal{E})$, by $\Theta \Lambda \Theta^{-1}$, leading to $\text{EVD}(\mathbf{I}_K + \mathcal{E}) = \Theta(\mathbf{I}_K + \Lambda)\Theta^{-1}$, i.e.

$$\det(\mathbf{I}_K + \mathcal{E}) = \prod_{i=1}^K (1 + \lambda_i) , \quad (3.80)$$

where the λ_i s are the eigenvalues of \mathcal{E} , assumed to be ordered by decreasing values. From the third order expansion of the logarithm function

$$\log(1 + \epsilon) = \epsilon - \frac{1}{2}\epsilon^2 + \frac{1}{3}\epsilon^3 + \dots , \quad (3.81)$$

we find using $\det(\mathbf{M}_1 \mathbf{M}_2) = \det \mathbf{M}_1 \det \mathbf{M}_2$ for $\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}(K)$

$$\begin{aligned}\log |\det((\mathbf{I}_K + \mathcal{E})\mathbf{B})| &= \log |\det \mathbf{B}| + \sum_{i=1}^K \log(1 + \lambda_i) \\ &= \log |\det \mathbf{B}| + \sum_{i=1}^K \lambda_i - \frac{1}{2} \sum_{i=1}^K \lambda_i^2 + \frac{1}{3} \sum_{i=1}^K \lambda_i^3 + \dots\end{aligned}$$

The above result assumes $|\lambda_i| < 1$ for all i , and this results from the hypothesis $\|\mathcal{E}\| \ll 1$.

As $\text{Tr}(\mathbf{M}_1 \mathbf{M}_2 \mathbf{M}_3) = \text{Tr}(\mathbf{M}_3 \mathbf{M}_1 \mathbf{M}_2)$ for $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3 \in \mathcal{M}(K)$, we have

$$\begin{aligned}\sum_{i=1}^K \lambda_i &= \text{Tr} \Lambda = \text{Tr}(\Lambda \underbrace{\Theta^{-1} \Theta}_{\mathbf{I}_K}) = \text{Tr}(\Theta \Lambda \Theta^{-1}) = \text{Tr} \mathcal{E} , \\ \sum_{i=1}^K \lambda_i^2 &= \text{Tr} \Lambda^2 = \text{Tr}(\Theta \Lambda^2 \Theta^{-1}) = \text{Tr}(\Theta \Lambda \underbrace{\Theta^{-1} \Theta}_{\mathbf{I}_K} \Lambda \Theta^{-1}) = \text{Tr} \mathcal{E}^2 , \\ \sum_{i=1}^K \lambda_i^3 &= \text{Tr} \Lambda^3 = \text{Tr}(\Theta \Lambda^3 \Theta^{-1}) = \text{Tr}(\Theta \Lambda \underbrace{\Theta^{-1} \Theta}_{\mathbf{I}_K} \Lambda \underbrace{\Theta^{-1} \Theta}_{\mathbf{I}_K} \Lambda \Theta^{-1}) = \text{Tr} \mathcal{E}^3 .\end{aligned}$$

Note that the eigenvalues of \mathcal{E} tend to zero when $\text{Tr} \mathcal{E}^2$ tends to zero as $\text{Tr} \mathcal{E}^2 = \sum_{i=1}^K |\lambda_i|^2$.

Finally, $o(\text{Tr} \mathcal{E}^2) = o(\|\mathcal{E}\|^2)$ (remind that for matrices, the norm symbol corresponds in this text to the Frobenius norm: $\|\mathcal{E}\|^2 = \sum_{ij} \mathcal{E}_{ij}^2$) because $\|\mathcal{E}\|^2$ does not tend faster to zero than $\text{Tr} \mathcal{E}^2$:

$$\begin{aligned}\|\mathcal{E}\|^2 - \text{Tr} \mathcal{E}^2 &= \sum_{i=1}^K \sum_{j=1}^K \mathcal{E}_{ij}^2 - \sum_{i=1}^K \sum_{j=1}^K \mathcal{E}_{ij} \mathcal{E}_{ji} \\ &= \sum_{i < j} \sum_{j < i} \mathcal{E}_{ij}^2 + \sum_{i < j} \sum_{j < i} \mathcal{E}_{ji}^2 - 2 \sum_{i < j} \sum_{j < i} \mathcal{E}_{ij} \mathcal{E}_{ji} \\ &= \sum_{i < j} \sum_{j < i} (\mathcal{E}_{ij} - \mathcal{E}_{ji})^2 \\ &\geq 0 .\end{aligned}\tag{3.82}$$

This means that a matrix \mathcal{E} having a small Frobenius norm will also have a small $\text{Tr} \mathcal{E}^2$. Another way to obtain this result is to observe that $\log |\det \mathbf{B}| + \text{Tr} \mathcal{E} - \frac{1}{2} \text{Tr} \mathcal{E}^2$ is an approximation of $\log |\det((\mathbf{I}_K + \mathcal{E})\mathbf{B})|$ up to a term of order $\sum_{i=1}^K \lambda_i^2$. But each of the eigenvalue (their modulus if some are complex) are bounded above by $\|\mathcal{E}\|^2$. Indeed, for the spectral norm (as well as all other p -norms), we have

$$\|\mathbf{M}\mathbf{x}\|_2 \leq \|\mathbf{M}\|_2 \|\mathbf{x}\|_2 .\tag{3.83}$$

In particular, setting \mathbf{x}_i the eigenvector of \mathbf{M} associated to the eigenvalue λ_i ($\|\mathbf{M}\mathbf{x}_i\|_2 = |\lambda_i| \cdot \|\mathbf{x}_i\|_2$), we have:

$$\begin{aligned} |\lambda_i| \cdot \|\mathbf{x}_i\|_2 &\leq \|\mathbf{M}\|_2 \|\mathbf{x}_i\|_2 \\ |\lambda_i| &\leq \|\mathbf{M}\|_2, \end{aligned} \quad (3.84)$$

i.e., $|\lambda_i| \leq \|\mathbf{M}\|$. Therefore, with $\mathcal{E} \leftarrow \mathbf{M}$: $\sum_{i=1}^K |\lambda_i|^2 \leq K \|\mathcal{E}\|^2$ and $O(\sum_{i=1}^K \lambda_i^2) = O(\|\mathcal{E}\|^2)$.

Hence, by taking a matrix \mathcal{E} with sufficiently small (Frobenius) norm, the following approximation holds:

$$\log |\det((\mathbf{I}_K + \mathcal{E})\mathbf{B})| \approx \log |\det \mathbf{B}| + \text{Tr} \mathcal{E} - \frac{1}{2} \text{Tr} \mathcal{E}^2. \quad (3.85)$$

□

3.8.2 Proof of Lemma 11 (wording p. 116)

Note that (S_1, S_2) is distributed as $(U_1 + \sigma Z_1, U_2 + \sigma Z_2)$ where U_1, U_2 are independent Bernoulli variables taking the value ± 1 with probability $1/2$ and Z_1, Z_2 are independent standard normal variables independent of U_1, U_2 . Thus (Y_1, Y_2) is distributed as $(U_1 + U_2 + \sqrt{2}\sigma Z'_1, U_2 - U_1 + \sqrt{2}\sigma Z'_2)$ where Z'_1, Z'_2 are also independent standard normal variables independent of U_1, U_2 . Since $U_1 + U_2$ and $U_1 - U_2$ both take the value ± 2 with probability $1/4$ and 0 with probability $1/2$, Y_1 and Y_2 have the same pdf as p_Y given in the lemma. Direct calculation yields $\psi_Y(y) = -p'_Y(y)/p_Y(y) = 1/(2\sigma^2) \sum_{i=-1}^1 (y - 2i) w_i(y)$ with w_i as defined in the lemma. Noting that

$$w'_i(y) = \sum_{j=-1}^1 \frac{y - 2j}{2\sigma^2} w_i(y) w_j(y) - \frac{y - 2i}{2\sigma^2} w_i(y)$$

and $1 - w_i = \sum_{j \neq i} w_j$, $(1/2\sigma^2) \sum_{i=-1}^1 (y - 2i) w'_i(y)$ equals

$$\begin{aligned} \sum_{-1 \leq i < j \leq 1} \frac{w_i(y) w_j(y)}{4\sigma^4} [2(y - 2i)(y - 2j) - (y - 2i)^2 - (y - 2j)^2] \\ = -\frac{1}{\sigma^4} \sum_{-1 \leq i < j \leq 1} (j - i)^2 w_i(y) w_j(y). \end{aligned}$$

This yields the expression for ψ'_Y given by Eq. (3.26) of the lemma.

Let us now compute $E[Y_2^2 | Y_1 = y]$. To that end, note that the conditional density of (U_1, U_2) given $Y_1 = y$ is

$$\begin{aligned} \Pr(U_1 = 1, U_2 = 1 | Y_1 = y) &= w_1(y), \\ \Pr(U_1 = 1, U_2 = -1 | Y_1 = y) &= w_0(y)/2, \\ \Pr(U_1 = -1, U_2 = 1 | Y_1 = y) &= w_0(y)/2, \\ \Pr(U_1 = -1, U_2 = -1 | Y_1 = y) &= w_{-1}(y). \end{aligned}$$

The above values are easily found using Bayes' formula

$$\Pr(U_1, U_2 | Y_1) = \frac{\Pr(Y_1 | U_1, U_2) \Pr(U_1, U_2)}{\Pr(Y_1)}. \quad (3.86)$$

Therefore, conditionally on $Y_1 = y$, $Y_2 = U_2 - U_1 + \sqrt{2}\sigma Z'_2$ is distributed as $\sqrt{2}\sigma Z'_2$ with probability $w_{-1}(y) + w_1(y)$ and as $\sqrt{2}\sigma Z'_2 \pm 2$ with probability $w_0(y)/2$ each. But

$$\begin{aligned} E[Y_2^2 | Y_1 = y] &= (w_{-1}(y) + w_1(y)) \underbrace{\int \xi^2 \frac{1}{\sqrt{2}\sigma} \phi\left(\xi/\sqrt{2}\sigma\right) d\xi}_{=\text{Var}[\sqrt{2}\sigma Z'_2]=2\sigma^2} \\ &\quad + \frac{w_0(y)}{2} \underbrace{\int \xi^2 \frac{1}{\sqrt{2}\sigma} \phi\left((\xi-2)/\sqrt{2}\sigma\right) d\xi}_{=\text{Var}[\sqrt{2}\sigma Z'_2]+E^2[\sqrt{2}\sigma Z'_2+2]} \\ &\quad + \frac{w_0(y)}{2} \underbrace{\int \xi^2 \frac{1}{\sqrt{2}\sigma} \phi\left((\xi+2)/\sqrt{2}\sigma\right) d\xi}_{=\text{Var}[\sqrt{2}\sigma Z'_2]+E^2[\sqrt{2}\sigma Z'_2-2]} \\ &= 2\sigma^2(w_{-1}(y) + w_1(y) + w_0(y)) + 4w_0(y), \end{aligned} \quad (3.87)$$

which is precisely the expression for $E[Y_2^2 | Y_1 = y]$ given in the lemma. \square

3.8.3 Proof of Lemma 12 (wording p. 116)

We have $E[Y_2^2 \psi'_{Y_1}(Y_1)] = \int E[Y_2^2 | Y_1 = y] \psi'_{Y_1}(y) p_Y(y) dy$. Hence noting that $w_0(y) = \phi[y/(\sqrt{2}\sigma)]/[2p_Y(y)]$, $w_{\pm 1}(y) = \phi[(y \mp 2)/(\sqrt{2}\sigma)]/[4p_Y(y)]$, and that w_0 is an even function, one gets the first result of the lemma.

We now derive an upper bound for $g(y) = [\sigma^2 + 2w_0(y)][w_0(y) + 2w_1(y)]/\sigma^4$. We have,

$$\frac{w_0(y)}{w_{-1}(y)} = 2e\left[\frac{(y+2)^2 - y^2}{4\sigma^2}\right] = 2e\left(\frac{y+1}{\sigma^2}\right).$$

Thus, since $w_{-1} = 1 - w_1 - w_0$

$$w_0(y) = \frac{2e^{(y+1)/\sigma^2}}{1 + 2e^{(y+1)/\sigma^2}} [1 - w_1(y)] \leq 2e^{(y+1)/\sigma^2}.$$

Similarly, $w_0(y)/w_1(y) = 2e\{[(y-2)^2 - y^2]/(4\sigma^2)\} = 2e[(1-y)/\sigma^2]$ and since $w_0 = 1 - w_1 - w_{-1}$:

$$w_1(y) = \frac{1 - w_{-1}(y)}{1 + 2e^{(1-x)/\sigma^2}} \leq \frac{1}{1 + 2e^{(1-x)/\sigma^2}} \leq \frac{e^{(y-1)/\sigma^2}}{2}$$

Thus, for $y \leq -1 - \xi$, one has

$$g(y) \leq (1 + 4e^{-\xi/\sigma^2}/\sigma^2)[2e^{-\xi/\sigma^2} + e^{-(2-\xi)/\sigma^2}]/\sigma^2$$

If we choose $\xi = \xi(\sigma)$ such that $\xi/\sigma^2 + \log \sigma^2 \rightarrow \infty$ and $\xi \rightarrow 0$ as $\sigma \rightarrow 0$, then both $e^{-\xi/\sigma^2}/\sigma^2$ and $e^{-(2-\xi)/\sigma^2}/\sigma^2$ tend to 0 as $\sigma \rightarrow 0$. Hence

$$\int_{-\infty}^{-1-\xi} g(y)\phi\left(\frac{y+2}{\sqrt{2}\sigma}\right)dy \rightarrow 0$$

For $x \geq -1 - \xi$, one can bound $w_0(y)$ and $w_1(y)$ by 1, hence $g(y)$ by $2(\sigma^2 + 2)$. Therefore

$$\int_{-1-\xi}^{\infty} g(y)\phi\left(\frac{y+2}{\sqrt{2}\sigma}\right)dy \leq \frac{3(\sigma^2 + 2)}{\sigma^4} \left[1 - \Phi\left(\frac{1-\xi}{\sqrt{2}\sigma}\right)\right]$$

where $\Phi(y) = \int_{-\infty}^y e^{-t^2/2}dt/\sqrt{2\pi}$ is the (cumulative) distribution function associated to the Gaussian density. But we know that $1 - \Phi(y) \leq e^{-y^2/2}/(y\sqrt{2\pi})$, hence

$$\frac{1}{\sigma^4} \left[1 - \Phi\left(\frac{1-\xi}{\sqrt{2}\sigma}\right)\right] \leq \frac{e^{-(1-\xi)^2/(4\sigma^2)}}{(1-\xi)\sqrt{\pi}\sigma^3} \rightarrow 0 \quad \text{as } \sigma \rightarrow 0,$$

since $\xi \rightarrow 0$ as $\sigma \rightarrow 0$.

□

3.8.4 Proof of Lemma 13 (wording p. 121)

We first prove that $\mathcal{H}[p]$ is an upper bound on $h(Y)$. We have from the definition of differential entropy that $h(Y) = \sum_{n=1}^N \pi_n H_n$ where

$$H_n \doteq - \int K_n(y) \log \left[\sum_{m=1}^N \pi_m K_m(y) \right] dy. \quad (3.88)$$

Since all $K_m \geq 0$, the last right hand side is bounded above by

$$- \int K_n(y) \log [\pi_n K_n(y)] dy = h[K_n] - \log \pi_n ,$$

yielding the first claim.

Inequality (3.31) can also easily be managed based on entropy properties. Indeed, the density given in Eq. (3.2) is a convex combination of densities. This mixture density has the nice property of being the marginal density of an augmented model (Y, U) , where U is a discrete variable with N values u_1, \dots, u_n and parameter $\boldsymbol{\pi}$ and $Y|U = u_n$ has density K_n for each n . The “continuous-discrete” joint entropy $h(Y, U)$ equals $h(Y|U) + H(U)$ where $H(U) \doteq -\sum_{n=1}^N \Pr(U = u_n) \log \Pr(U = u_n) = H[\boldsymbol{\pi}]$ and $h(Y|U) = E_U[h(Y|U = n)]$ is the equivocation

[M. Hellman, 1970] and reduces to $\sum_{n=1}^N \pi_n h[K_n]$. But on the other hand, we have $h(Y) = h(Y, U) - H(U|Y)$ and thus $H[p] - h[p]$ equals $H(U|Y)$ which is always nonnegative because U is a discrete variable.

Yet another way to prove (3.31) is to use the so-called generalized Jensen-Shannon (GJS) divergence. This generalized divergence is an extension of the well-known Jensen-Shannon divergence between two densities. The GJS divergence between the K_n is here defined as [Lin, 1991]

$$JS_{\boldsymbol{\pi}}(K_1, \dots, K_N) \doteq h \left[\sum_{n=1}^N \pi_n K_n \right] - \sum_{n=1}^N \pi_n h[K_n] . \quad (3.89)$$

Based on a result of Hellman and Raviv [M. Hellman, 1970], Lin showed in [Lin, 1991] using simple entropy properties that half the difference between $H[\boldsymbol{\pi}]$ and $JS_{\boldsymbol{\pi}}(K_1, \dots, K_N)$ is an upper bound on a positive term. Therefore,

$$h \left[\sum_{n=1}^N \pi_n K_n \right] - \sum_{n=1}^N \pi_n h[K_n] \leq H[\boldsymbol{\pi}] \quad (3.90)$$

which shows that $h[p] - H[p] \leq 0$.

Let us now turn to the lower bound of the entropy. To prove the inequality (3.32), note that $\log(1+x) \leq x$. The term $\log[\sum_{m=1}^N \pi_m K_m(y)]$ can be bounded above by

$$\begin{cases} \log[\pi_n K_n(y)] + \sum_{1 \leq m \leq N, m \neq n} \frac{\pi_m K_m(y)}{\pi_n K_n(y)} & \text{if } y \in \Omega_n \\ \log(\max_{1 \leq m \leq N} \sup K_m) & \text{otherwise} . \end{cases} \quad (3.91)$$

Therefore, from (3.88), one gets

$$\begin{aligned} H_n &\geq - \int_{\Omega_n} K_n(y) \log[\pi_n K_n(y)] dy - \sum_{1 \leq m \leq N, m \neq n} \frac{\pi_m}{\pi_n} \int_{\Omega_n} K_m(y) dy \\ &\quad - \log(\max_{1 \leq m \leq N} \sup K_m) \epsilon_n . \end{aligned}$$

But since the subsets $\Omega_1, \dots, \Omega_N$ are disjoint,

$$\sum_{n=1}^N \pi_n \sum_{1 \leq m \leq N, m \neq n} \frac{\pi_m}{\pi_n} \int_{\Omega_n} K_m(y) dy = \sum_{m=1}^N \pi_m \int_{\cup_{1 \leq n \neq m \leq N} \Omega_n} K_m(y) dy ,$$

and $\cup_{1 \leq n \neq m \leq N} \Omega_n \subseteq \mathbb{R} \setminus \Omega_m$. Therefore the right hand side of the above equality is bounded above by $\sum_{m=1}^N \pi_m \epsilon_m$. Hence,

$$\begin{aligned}
\sum_{n=1}^N \pi_n H_n &\geq - \sum_{n=1}^N \pi_n \int_{\Omega_n} K_n(y) \log \pi_n dy - \sum_{n=1}^N \pi_n \int_{\Omega_n} K_n(y) \log K_n(y) dy \\
&\quad - \sum_{n=1}^N \pi_n \epsilon_n - \sum_{n=1}^N \pi_n \log(\max_{1 \leq m \leq N} \sup K_m) \epsilon_n \\
&= - \sum_{n=1}^N \pi_n \log \pi_n (1 - \epsilon_n) \\
&\quad - \sum_{n=1}^N \pi_n \left[-h[K_n] - \int_{\mathbb{R} \setminus \Omega_n} K_n(y) \log K_n(y) dy \right] \\
&\quad - \sum_{n=1}^N \pi_n \epsilon_n - \sum_{n=1}^N \pi_n \log(\max_{1 \leq m \leq N} \sup K_m) \epsilon_n .
\end{aligned}$$

It follows that $h[p] = \sum_{n=1}^N \pi_n H_n$ is bounded below by

$$\begin{aligned}
H[\pi] + \sum_{n=1}^N \pi_n h[K_n] + \sum_{n=1}^N \pi_n \log(\pi_n \sup K_n) \epsilon_n - \sum_{n=1}^N \pi_n \epsilon'_n \\
- \sum_{m=1}^N \pi_m \epsilon_m - \sum_{n=1}^N \pi_n \log(\max_{1 \leq m \leq N} \sup K_m) \epsilon_n .
\end{aligned}$$

and final basic manipulations yield the lower bound given in the lemma. \square

3.8.5 Proof of Lemma 14 (wording p. 126)

By construction, for each $j = 1, \dots, r$, $\mathbf{w}^* \mathbf{u}$ takes the same values for $\mathbf{u} \in \mathcal{U}_j$. On the other hand, by grouping the vectors $\mathbf{u} \in \mathcal{U}$ which produce the same value of $\mathbf{w}^* \mathbf{u}$ into subsets of \mathcal{U} , one gets a partition of \mathcal{U} into $r^* + 1$ subsets $\mathcal{U}_0^*, \dots, \mathcal{U}_{r^*}^*$, such that each $\mathcal{U}_j^*, 1 \leq j \leq r^*$ contains at least two elements and $\mathbf{w}^* \mathbf{u}$ takes the same values for $\mathbf{u} \in \mathcal{U}_j^*$ and the values associated with different \mathcal{U}_j^* and the $\mathbf{w}^* \mathbf{u}, \mathbf{u} \in \mathcal{U}_0^*$, are all distinct. Obviously $r^* \geq 1$ and each of the $\mathcal{U}_1, \dots, \mathcal{U}_r$, must be contained in one of the $\mathcal{U}_1^*, \dots, \mathcal{U}_{r^*}^*$. Therefore the space V must be contained in the space spanned by the vectors $\mathbf{u} - \mathbf{u}_j, \mathbf{u} \in \mathcal{U}_j^* \setminus \{\mathbf{u}_j\}$, $j = 1, \dots, r^*$, and $\mathbf{u}_1, \dots, \mathbf{u}_{r^*}$ being arbitrary elements of $\mathcal{U}_1^*, \dots, \mathcal{U}_{r^*}^*$. But the last space is orthogonal to \mathbf{w}^* by construction and thus cannot have dimension greater than $K - 1$, hence it must coincide with V .

Putting $\Pr(\mathbf{u})$ for $\Pr(\mathbf{U} = \mathbf{u})$ for short and $\Pr(\mathcal{U}_j^*) = \sum_{\mathbf{u} \in \mathcal{U}_j^*} \Pr(\mathbf{u})$, one has

$$H(\mathbf{w}^* \mathbf{U}) = - \sum_{\mathbf{u} \in \mathcal{U}_0^*} \Pr(\mathbf{u}) \log \Pr(\mathbf{u}) - \sum_{j=1}^{r^*} \Pr(\mathcal{U}_j^*) \log \Pr(\mathcal{U}_j^*) .$$

For a given pair \mathbf{u}, \mathbf{u}' of distinct vectors in \mathcal{U} , if $\mathbf{w}^*(\mathbf{u} - \mathbf{u}') \neq 0$ then it remains so when \mathbf{w}^* is changed to \mathbf{w} provided that the change is sufficiently small. But if $\mathbf{w}^*(\mathbf{u} - \mathbf{u}') = 0$ then this equality may break however small the change. In fact if \mathbf{w} is not proportional to \mathbf{w}^* , it is not orthogonal to V , hence $\mathbf{w}(\mathbf{u} - \mathbf{u}') \neq 0$ for at least one pair \mathbf{u}, \mathbf{u}' of distinct points in some \mathcal{U}_j^* , meaning that $\mathbf{w}\mathbf{u}$ takes at least two distinct values in \mathcal{U}_j^* . Thus there exists a neighborhood \mathcal{W} of \mathbf{w}^* in \mathcal{S} such that for all $\mathbf{w} \in \mathcal{W} \setminus \{\mathbf{w}^*\}$, each subset \mathcal{U}_j^* can be partitioned into subsets $\mathcal{U}_{j,k}(\mathbf{w}), k = 1, \dots, n_j(\mathbf{w})$ ($n_j(\mathbf{w})$ can be 1) such that $\mathbf{w}\mathbf{u}$ takes the same value on $\mathcal{U}_{j,k}(\mathbf{w})$, and the values of $\mathbf{w}\mathbf{u}$ on the subsets $\mathcal{U}_{j,k}(\mathbf{w})$ and on each point of \mathcal{U}_0^* are distinct. Further, there exists at least one index i for which $n_i(\mathbf{w}) > 1$. For such an index

$$\begin{aligned} \Pr(\mathcal{U}_i^*) \log \Pr(\mathcal{U}_i^*) &= \sum_{k=1}^{n_i(\mathbf{w})} \Pr(\mathcal{U}_{i,k}(\mathbf{w})) \log \Pr(\mathcal{U}_{i,k}(\mathbf{w})) \\ &\quad + \sum_{k=1}^{n_i(\mathbf{w})} \Pr(\mathcal{U}_{i,k}(\mathbf{w})) \log \frac{\Pr(\mathcal{U}_i^*)}{\Pr(\mathcal{U}_{i,k}(\mathbf{w}))}. \end{aligned}$$

The last term can be seen to be a strictly positive number, as $\Pr(\mathcal{U}_i^*) > \Pr(\mathcal{U}_{i,k}(\mathbf{w}))$ for $1 \leq k \leq n_i(\mathbf{w})$ once $n_i(\mathbf{w}) > 1$. Note that this term does not depend directly on \mathbf{w} but only indirectly via the set $\mathcal{U}_{j,k}(\mathbf{w}), k = 1, \dots, n_j(\mathbf{w}), j = 1, \dots, r^*$, and there is only a finite number of possible such sets. Therefore $H(\mathbf{w}\mathbf{U}) \geq H(\mathbf{w}^*\mathbf{U}) + \alpha$ for some $\alpha > 0$ for all $\mathbf{w} \in \mathcal{W}$.

In the $K = 2$ case, the space V reduces to a line and thus the differences $\mathbf{u} - \mathbf{u}'$ for distinct \mathbf{u}, \mathbf{u}' in \mathcal{U}_j^* , for all j , belong to this line. Thus if \mathbf{w} is not proportional to \mathbf{w}^* , hence not orthogonal to this line, $\mathbf{w}\mathbf{u}$ takes distinct values on each of the sets $\mathcal{U}_1^*, \dots, \mathcal{U}_{r^*}^*$, and if \mathbf{w} is close enough to \mathbf{w}^* , these values are also distinct for different sets and distinct from the values of $\mathbf{w}\mathbf{u}$ on \mathcal{U}_0^* , which are distinct themselves. Thus for such \mathbf{w} , $H(\mathbf{w}\mathbf{U}) = H(\mathbf{U})$.

□

3.8.6 Proof of Lemma 15 (wording p. 130)

The proof of this lemma is quite involved in the $K > 2$ case, therefore, we will first give the proof for the $K = 2$ case which is much simpler, and then proceed by extending it to $K > 2$. As already shown in the beginning of Section 3.2.3.2, $\mathbf{w}\mathbf{S} = \mathbf{w}\mathbf{U} + \sigma Z$ where Z is a standard Gaussian random variable. Thus, the density of $\mathbf{w}\mathbf{S}$ is of the form (3.2) with $K_n(y) = \phi[(y - \mu_n)/\sigma]/\sigma$, μ_1, \dots, μ_N being the possible values of $H(\mathbf{w}\mathbf{U})$ and ϕ being the standard Gaussian density, as usual. For $\mathbf{w} = \mathbf{w}^*$, one has by Lemma 13,

$$h(\mathbf{w}^*\mathbf{S}) \leq H(\mathbf{w}^*\mathbf{U}) + h[\phi] + \log \sigma.$$

On the other hand, we have seen in the proof of Lemma 14 that for \mathbf{w} in some neighborhood \mathcal{W} of \mathbf{w}^* and distinct from \mathbf{w} , the $\mathbf{w}\mathbf{u}, \mathbf{u} \in \mathcal{U}$ (\mathcal{U} denoting

the set of possible values of U) are all distinct (in the $K = 2$ case). Thus the maps $\mathbf{u} \mapsto \mathbf{w}\mathbf{u}$ map different points $\mathbf{u} \in \mathcal{U}$ to different μ_n . However, when \mathbf{w} approaches \mathbf{w}^* , some of the μ_n tend to coincide and thus some of the d_n defined in (3.36) approach zero. To avoid this we restrict \mathbf{w} to $\mathcal{W} \setminus \mathcal{W}'$ where \mathcal{W}' is any open neighborhood of \mathbf{w}^* strictly included in \mathcal{W} . Then $\min_n d_n \geq d$ for all $\mathbf{w} \in \mathcal{W} \setminus \mathcal{W}'$ for some $d > 0$ (which depends on \mathcal{W}'). Thus by Corollary 12, $h(\mathbf{w}\mathbf{S})$ can be made arbitrarily close to $H(\mathbf{w}U) + h[\phi] + \log \sigma$ for all $\mathbf{w} \in \mathcal{W} \setminus \mathcal{W}'$ by taking σ small enough. But $H(\mathbf{w}U) = H(U) > H(\mathbf{w}^*U)$, therefore $h(\mathbf{w}\mathbf{S}) > h(\mathbf{w}^*\mathbf{S})$ for all $\mathbf{w} \in \mathcal{W} \setminus \mathcal{W}'$, for σ small enough.

One can always choose \mathcal{W} to be a closed set in $\mathcal{S}(K)$; hence it is compact. Since the function $\mathbf{w} \in \mathcal{W} \mapsto h(\mathbf{w}\mathbf{S})$ is continuous, it must admit a minimum, which by the above result must be in \mathcal{W}' and thus is not on the boundary of \mathcal{W} . This shows that this minimum is a local minimum. Finally, as one can choose \mathcal{W}' arbitrarily small, the above result shows that the above local minimum converges to \mathbf{w}^* as $\sigma \rightarrow 0$.

Consider now the case $K > 2$. The difficulty is that it is no longer true that for \mathbf{w} in some neighborhood \mathcal{W} of \mathbf{w}^* and distinct from \mathbf{w}^* , the $\mathbf{w}\mathbf{u}, \mathbf{u} \in \mathcal{U}$ are all distinct. Indeed, by construction of \mathbf{w}^* , there exists $K - 1$ pairs $(\mathbf{u}_j, \mathbf{u}'_j), 1 \leq j < K$, of distinct vectors in \mathcal{U} such that the differences $\mathbf{u}_j - \mathbf{u}'_j$ are linearly independent and $\mathbf{w}^*(\mathbf{u}_j - \mathbf{u}'_j) = 0, 1 \leq j < K$ (for a given \mathbf{w}^*). For \mathbf{w} not proportional to \mathbf{w}^* , at least one (but not necessarily all) of the above equalities will break. Therefore all the $\mathbf{w}\mathbf{u}, \mathbf{u} \in \mathcal{U}$ may be not distinct, even if \mathbf{w} is restricted to $\mathcal{W} \setminus \mathcal{W}'$. But the set of \mathbf{w} for which this property is not true anymore is the union of a finite number of linear subspaces of dimension $K - 1$ of \mathbb{R}^K and thus is not dense in \mathbb{R}^K . Therefore for most of the $\mathbf{w} \in \mathcal{W} \setminus \mathcal{W}'$, the $\mathbf{w}\mathbf{u}, \mathbf{u} \in \mathcal{U}$ are all distinct.

The pdf of $\mathbf{w}\mathbf{S}$ can be written as

$$p(y) = \sum_{\mathbf{u} \in \mathcal{U}} \Pr(\mathbf{u}) \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{w}\mathbf{u}}{\sigma}\right); \quad (3.92)$$

but some of the $\mathbf{w}\mathbf{u}, \mathbf{u} \in \mathcal{U}$ can be arbitrarily close to each other. In this case it is of interest to group the corresponding terms in (3.92) together. Thus we rewrite $p(y)$ as

$$p(y) = \sum_{n=1}^N \sum_{\mathbf{u} \in \mathcal{V}_n} \Pr(\mathbf{u}) \left[\sum_{\mathbf{u} \in \mathcal{V}_n} \frac{\Pr(\mathbf{u})}{\sum_{\mathbf{u} \in \mathcal{V}_n} \Pr(\mathbf{u})} \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{w}\mathbf{u}}{\sigma}\right) \right],$$

where $\mathcal{V}_1, \dots, \mathcal{V}_N$ is a partition of \mathcal{U} . This pdf is still of the form (3.2) with

$$\pi_n = \sum_{\mathbf{u} \in \mathcal{V}_n} \Pr(\mathbf{u}), \quad K_n(y) = \sum_{\mathbf{u} \in \mathcal{V}_n} \frac{\Pr(\mathbf{u})}{\pi_n} \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{w}\mathbf{u}}{\sigma}\right).$$

The partition $\mathcal{V}_1, \dots, \mathcal{V}_N$ can and should be chosen so that

$$d(\mathbf{w}) \doteq \min_{1 \leq n \neq m \leq N} \min_{\mathbf{u} \in \mathcal{V}_n, \mathbf{u}' \in \mathcal{V}_m} |\mathbf{w}\mathbf{u} - \mathbf{w}\mathbf{u}'|,$$

is bounded below by some given positive number. In other words, all the vectors $\mathbf{u} \in \mathcal{U}$ yielding the same value of $\mathbf{w}\mathbf{u}$ must be grouped in a same subset \mathcal{N}_n . To this end, note that, as is shown in the proof of Lemma 14, \mathbf{w}^* is associated with a partition $\mathcal{U}_0^*, \dots, \mathcal{U}_r^*$ of \mathcal{U} such that $\mathbf{w}^*\mathbf{u}$ takes the same value for all $\mathbf{u} \in \mathcal{U}_j^*$ ($1 \leq j \leq r^*$), and the values associated with different \mathcal{U}_j^* and the $\mathbf{w}^*\mathbf{u}, \mathbf{u} \in \mathcal{U}_0^*$, are all distinct. Thus $\inf_{\mathbf{w} \in \mathcal{W}} |\mathbf{w}\mathbf{u} - \mathbf{w}\mathbf{u}'| \geq \delta$ for some $\delta > 0$ for all $\mathbf{u} \neq \mathbf{u}'$ and \mathbf{u}, \mathbf{u}' do not belong to a same $\mathcal{U}_j^*, j = 1, \dots, r^*$. We take $N = r^* + \#\mathcal{U}_0^*$, where $\#\mathcal{U}_0^*$ denotes the number of elements of \mathcal{U}_0^* , $\mathcal{V}_j = \mathcal{U}_j^*, j = 1, \dots, r^*$ and the remaining \mathcal{V}_j to be disjoint sets containing only a single element of \mathcal{U}_0^* . Then, the partition $\{\mathcal{V}_1, \dots, \mathcal{V}_N\} = \{\{\mathbf{u}\}, \mathbf{u} \in \mathcal{U}_0^*, \mathcal{U}_1^*, \dots, \mathcal{U}_{r^*}^*\}$ satisfies $d(\mathbf{w}) \geq \delta, \forall \mathbf{w} \in \mathcal{W}$. The above partition is not fine enough in order to apply Lemma 13 and to obtain the desired lower bound for $H[p]$. The application of this lemma with $\pi_n, K_n, n = 1, \dots, N$ would yield a lower bound involving $H[\boldsymbol{\pi}]$. By construction, $H[\boldsymbol{\pi}] = H[\mathbf{w}^*\mathbf{U}]$ while we would need a strict inequality. We thus refine the partition $\{\mathcal{V}_1, \dots, \mathcal{V}_N\}$ by splitting one of the sets $\mathcal{U}_j^*, j = 1, \dots, r^*$ into two subsets. The splitting rule is as follows: for each \mathcal{U}_j^* arrange the $\mathbf{w}\mathbf{u}, \mathbf{u} \in \mathcal{U}_j^*$ in ascending order and look for the maximum gap between two consecutive values. The set \mathcal{U}_j^* that produces the largest gap will be split and the splitting is done at the gap. For $\mathbf{w} \in \mathcal{W} \setminus \mathcal{W}'$, this maximum gap can be bounded below by a positive number δ' (noting that there is only a finite number of elements in each \mathcal{U}_j^*); hence for the refined partition, $d(\mathbf{w}) \geq \min(\delta, \delta')$. Of course, the partition constructed this way depends on \mathbf{w} , but there can be only a finite number of possible partitions. Hence, one can find a finite number of subsets $\mathcal{W}_1, \dots, \mathcal{W}_q$ which cover $\mathcal{W} \setminus \mathcal{W}'$, each of which is associated with a partition of \mathcal{U} such that the corresponding $d(\mathbf{w})$ is bounded below by $\min(\delta, \delta')$ for all \mathbf{w} in this subset. In the following we shall restrict \mathbf{w} to one such subset, \mathcal{W}_p say, and we denote by $\mathcal{V}_1, \dots, \mathcal{V}_N$ the associated partition.⁴

We now apply Lemma 13 with $\pi_n, K_n, n = 1, \dots, N$ defined as above and with the sets Ω_n defined by

$$\Omega_n \doteq \{y : \min_{\mathbf{u} \in \mathcal{V}_n} |y - \mathbf{w}\mathbf{u}| < d(\mathbf{w})/2\}.$$

Then we have, writing d in place of $d(\mathbf{w})$ for short,

$$\begin{aligned} \epsilon_n &\leq 1 - \int_{-d/(2\sigma)}^{d/(2\sigma)} \phi(x)dx = \text{Erfc}\left(\frac{d}{2\sqrt{2}\sigma}\right) \\ \epsilon'_n &= \sum_{\mathbf{u} \in \mathcal{V}_n} \frac{\Pr(\mathbf{u})}{\pi_n} \int_{\mathbb{R} \setminus \Omega_n} \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{w}\mathbf{u}}{\sigma}\right) \log \frac{\sup K_n}{K_n(y)} dy. \end{aligned}$$

In each term in the sum in that last right hand side, one applies the bound

$$\frac{\sup K_n}{K_n(y)} \leq \frac{\sigma \sup K_n}{[\Pr(\mathbf{u})/\pi_n] \phi[(y - \mathbf{w}\mathbf{u})/\sigma]}$$

⁴Note that the partition obtained *after the split* obviously counts one more element than the corresponding partition *before the split*. However, the same symbol N is used for both partitions to simplify the notation

which yields,

$$\begin{aligned}\epsilon'_n &\leq \sum_{\mathbf{u} \in \mathcal{V}_n} \frac{\Pr(\mathbf{u})}{\pi_n} \int_{|x|>d/(2\sigma)} \phi(x) \log \frac{\sigma \sup K_n}{[\Pr(\mathbf{u})/\pi_n] \phi(x)} dx \\ &= \left[\log \sup(\sigma K_n) - \sum_{\mathbf{u} \in \mathcal{V}_n} \frac{\Pr(\mathbf{u})}{\pi_n} \log \frac{\Pr(\mathbf{u})}{\pi_n} \right] \operatorname{Erfc}\left(\frac{d}{2\sqrt{2}\sigma}\right) \\ &\quad + h[\phi] - H_{d/\sigma}(\phi).\end{aligned}$$

Therefore, putting $h_n = -\sum_{\mathbf{u} \in \mathcal{V}_n} [\Pr(\mathbf{u})/\pi_n] \log [\Pr(\mathbf{u})/\pi_n]$ and noting further that $\sup(\sigma K_n) \leq \sup \phi = (2\pi)^{-1/2}$, one gets

$$\begin{aligned}\sum_{n=1}^N \pi_n \epsilon'_n + \sum_{n=1}^N \pi_n \left[\log \left(\frac{\max_{1 \leq m \leq N} \sup K_m}{\pi_n \sup K_n} \right) + 1 \right] \epsilon_n &\leq \\ \left[1 - \frac{\log(2\pi)}{2} + \sum_{n=1}^N \pi_n h_n \right] \operatorname{Erfc}\left(\frac{d}{2\sqrt{2}\sigma}\right) + h[\phi] - H_{d/\sigma}(\phi) &.\end{aligned}$$

Since $d = d(\mathbf{w}) \geq \min(\delta, \delta')$, $\forall \mathbf{w} \in \mathcal{W}_p$, the last inequality shows that for any $\eta > 0$,

$$h[p] \geq \sum_{n=1}^N \pi_n h[K_n] + H[\boldsymbol{\pi}] - \eta, \quad \forall \mathbf{w} \in \mathcal{W}_p,$$

for σ small enough. On the other hand, since $\log x \leq x - 1$,

$$\int \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{w}\mathbf{u}}{\sigma}\right) \log \frac{K_n(y)}{\phi[(y - \mathbf{w}\mathbf{u})/\sigma]/\sigma} dy \leq 0.$$

Multiplying both members of the above inequality by $\Pr(\mathbf{u})/\pi_n$ and summing up with respect to $\mathbf{u} \in \mathcal{V}_n$, one gets $h[\phi] + \log \sigma - h[K_n] \leq 0$. Therefore

$$h[p] \geq h[\phi] + \log \sigma + H[\boldsymbol{\pi}] - \eta.$$

But by construction $H[\boldsymbol{\pi}] > H(\mathbf{w}^* \mathbf{U})$ (see the proof of Lemma 14); therefore, taking $\eta < H[\boldsymbol{\pi}] - H(\mathbf{w}^* \mathbf{U})$, one sees that for σ small enough $h(\mathbf{w}\mathbf{S}) = h[p] > h(\mathbf{w}^* \mathbf{S})$ for all $\mathbf{w} \in \mathcal{W}_p$. Since this is true for all $p = 1, \dots, q$, we conclude as before that $h(\mathbf{w}\mathbf{S})$ admits a local minimum in \mathcal{W}' .

□

3.8.7 Proof of Lemma 16 (wording p. 140)

Let us fix the distinct indexes $1 \leq i, j \leq K$ and the small scalar ζ . Note that in some pathological cases, the sign of ζ cannot be arbitrarily chosen, otherwise, the $\mathbf{w} + \delta \mathbf{w}_{ij}^\zeta \in \mathcal{V}_K^1$ may be not satisfied (for example, if $\mathbf{w} = \mathbf{e}_i$, then we must obviously take $\zeta < 0$ and $\xi > 0$). The $\mathbf{w} + \delta \mathbf{w}_{ij}^\zeta \in \mathcal{V}_K^1$ constraint yields:

$$\xi^2 + 2\mathbf{w}(j)\xi + \zeta^2 + 2\mathbf{w}(i)\zeta = 0. \quad (3.93)$$

Both roots of the previous equation will lead to the same absolute value of $\mathbf{w}(r) + \delta\mathbf{w}_{ij}^\zeta(r)$, for all $1 \leq r \leq K$. We focus on the single root of (3.93) satisfying $|\xi| < |\mathbf{w}(j)|$ ($\mathbf{w} + \delta\mathbf{w}_{ij}^\zeta \in \mathcal{V}_K^1$), which gives (3.42).

With this value of ξ , observe that $\|\delta\mathbf{w}_{ij}^\zeta\| \rightarrow 0$ as $|\zeta| \rightarrow 0^5$.

Finally, by definition of $\delta\mathbf{w}_{ij}^\zeta$, the r -th entry of \mathbf{w} equals the r -th entry of $\mathbf{w} + \delta\mathbf{w}_{ij}^\zeta$ except if $r \in \{i, j\}$, which gives the $\Delta\tilde{\mathcal{C}}_R(\mathbf{w} + \delta\mathbf{w}_{ij}^\zeta, \mathbf{w})$ given in the lemma.

□

3.8.8 Proof of Theorem 18 (wording p. 141)

We freely assume $\zeta > 0$. If $\Delta\tilde{\mathcal{C}}_R^1 > 0$, the Theorem is obviously trivially proven. Consider then the unique alternative $\Delta\tilde{\mathcal{C}}_R^1 \leq 0$; we will show that in this case, $\Delta\tilde{\mathcal{C}}_R^2 > 0$.

Combination of equations (3.42) and (3.43) with $\mathbf{w} \notin \{\mathbf{e}_1, \dots, \mathbf{e}_K\}$, $\Delta\tilde{\mathcal{C}}_R^1 \leq 0$ and ζ a strictly positive small scalar gives:

$$\begin{aligned} -\Delta\tilde{\mathcal{C}}_R^1 &= R(\mathbf{S}_i)\zeta + R(\mathbf{S}_j) \left(-\mathbf{w}(j) + \sqrt{\mathbf{w}(j)^2 - (2\mathbf{w}(i)\zeta + \zeta^2)} \right) \\ &\geq 0 . \end{aligned} \quad (3.94)$$

Then,

$$R(\mathbf{S}_i)\zeta \geq R(\mathbf{S}_j) \left(\mathbf{w}(j) - \sqrt{\mathbf{w}(j)^2 - (2\mathbf{w}(i)\zeta + \zeta^2)} \right) . \quad (3.95)$$

On the other hand,

$$-\Delta\tilde{\mathcal{C}}_R^2 = R(\mathbf{S}_i)(-\zeta) + R(\mathbf{S}_j) \left(-\mathbf{w}(j) + \sqrt{\mathbf{w}(j)^2 - (-2\mathbf{w}(i)\zeta + \zeta^2)} \right) ,$$

i.e.

$$R(\mathbf{S}_i)\zeta = \Delta\tilde{\mathcal{C}}_R^2 - R(\mathbf{S}_j)\mathbf{w}(j) + R(\mathbf{S}_j)\sqrt{\mathbf{w}(j)^2 - (-2\mathbf{w}(i)\zeta + \zeta^2)} .$$

Hence, by Eq. (3.95):

$$\Delta\tilde{\mathcal{C}}_R^2 - R(\mathbf{S}_j) \left(\mathbf{w}(j) - \sqrt{\mathbf{w}(j)^2 - (-2\mathbf{w}(i)\zeta + \zeta^2)} \right)$$

is greater than or equal to

$$R(\mathbf{S}_j) \left(\mathbf{w}(j) - \sqrt{\mathbf{w}(j)^2 - (2\mathbf{w}(i)\zeta + \zeta^2)} \right) ,$$

⁵Observe that there is no restriction to make ζ tending to zero since \mathcal{V}_K^1 is a connected set: \mathcal{V}_K^λ defines the surface of the K -dimensional sphere centered at the origin with radius λ in R_+^K , i.e. a continuous manifold in R_+^K .

yielding

$$\begin{aligned}\Delta \tilde{\mathcal{C}}_R^2 &\geq R(\mathbf{S}_j) \left(2\mathbf{w}(j) - \sqrt{\mathbf{w}(j)^2 - (2\mathbf{w}(i)\zeta + \zeta^2)} - \sqrt{\mathbf{w}(j)^2 - (-2\mathbf{w}(i)\zeta + \zeta^2)} \right) \\ &\geq R(\mathbf{S}_j)\mathbf{w}(j) \left(\left[1 - \sqrt{1 - \frac{2\mathbf{w}(i)\zeta + \zeta^2}{\mathbf{w}(j)^2}} \right] + \left[1 - \sqrt{1 + \frac{2\mathbf{w}(i)\zeta - \zeta^2}{\mathbf{w}(j)^2}} \right] \right) .\end{aligned}$$

Then, using Taylor development,

$$\begin{cases} \sqrt{1-\epsilon} &= 1 - \frac{\epsilon}{2} - \frac{\epsilon^2}{8} + o(\epsilon^2) \\ \sqrt{1+\epsilon'} &= 1 + \frac{\epsilon'}{2} - \frac{\epsilon'^2}{8} + o(\epsilon'^2) \end{cases}, \quad (3.96)$$

where $o(\epsilon^2)$ and $o(\epsilon'^2)$ denote terms tending to zero faster than $\|\epsilon^2\|$ and $\|\epsilon'^2\|$, respectively. Hence, for sufficiently small ϵ, ϵ' , one gets:

$$\begin{cases} 1 - \sqrt{1-\epsilon} &> 1 - (1 - \frac{\epsilon}{2}) \\ 1 - \sqrt{1+\epsilon'} &> 1 - (1 + \frac{\epsilon'}{2}) \end{cases}. \quad (3.97)$$

Then, by letting $\epsilon \doteq \zeta \frac{2\mathbf{w}(i)+\zeta}{\mathbf{w}(j)^2}$ and $\epsilon' \doteq \zeta \frac{2\mathbf{w}(i)-\zeta}{\mathbf{w}(j)^2}$, we have for ζ small enough:

$$\begin{cases} 1 - \sqrt{1 - \frac{2\mathbf{w}(i)\zeta + \zeta^2}{\mathbf{w}(j)^2}} &> 1 - (1 - \frac{2\mathbf{w}(i)\zeta + \zeta^2}{2\mathbf{w}(j)^2}) \\ 1 - \sqrt{1 + \frac{2\mathbf{w}(i)\zeta - \zeta^2}{\mathbf{w}(j)^2}} &> 1 - (1 + \frac{2\mathbf{w}(i)\zeta - \zeta^2}{2\mathbf{w}(j)^2}) \end{cases}. \quad (3.98)$$

By (3.98) and using inequality (3.96), it comes that for sufficiently small $\zeta > 0$:

$$\begin{aligned}\Delta \tilde{\mathcal{C}}_R^2 &> R(\mathbf{S}_j)\mathbf{w}(j) \left(\frac{2\mathbf{w}(i)\zeta + \zeta^2}{2\mathbf{w}(j)^2} - \frac{2\mathbf{w}(i)\zeta - \zeta^2}{2\mathbf{w}(j)^2} \right) \\ &= \frac{R(\mathbf{S}_j)\zeta^2}{\mathbf{w}(j)} \\ &> 0.\end{aligned}$$

□

3.8.9 Proof of Lemma 17 (wording p. 149)

The computation of the first derivative is trivial; it yields the stationary point condition:

$$\frac{R(\mathbf{S}_i)}{|w_i|} = \frac{\sum_{l=1}^K |w_l| R(\mathbf{S}_l)}{\|\mathbf{w}\|^2}. \quad (3.99)$$

Let us now turn to the second derivative expressions at stationary points. For $i \neq j$:

$$\begin{aligned}
\frac{\partial^2 \frac{R(\mathbf{w}\mathbf{S})}{\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]}}}{\partial w_i \partial w_j} &= \frac{(\text{sign}(w_i)R(\mathbf{S}_i)2w_j - w_i \text{sign}(w_j)R(\mathbf{S}_j)) \|\mathbf{w}\|^2}{\|\mathbf{w}\|^5} \\
&\quad - \frac{(\text{sign}(w_i)R(\mathbf{S}_i)\|\mathbf{w}\|^2 - \sum_{l=1}^K |w_l|R(\mathbf{S}_l)) 3w_j \|\mathbf{w}\|}{\|\mathbf{w}\|^5} \\
&= -w_j \text{sign}(w_i)R(\mathbf{S}_i)\|\mathbf{w}\|^{-3} - w_i \text{sign}(w_j)R(\mathbf{S}_j)\|\mathbf{w}\|^{-3} \\
&\quad + 3w_i w_j \sum_{l=1}^K |w_l|R(\mathbf{S}_l)\|\mathbf{w}\|^{-5} \\
&= (-w_j \text{sign}(w_i)R(\mathbf{S}_i) + 2w_i \text{sign}(w_j)R(\mathbf{S}_j))\|\mathbf{w}\|^{-3} \\
&= \frac{w_i w_j \sum_{l=1}^K |w_l|R(\mathbf{S}_l)}{\|\mathbf{w}\|^5}. \tag{3.100}
\end{aligned}$$

On the other hand

$$\begin{aligned}
\frac{\partial^2 \frac{R(\mathbf{w}\mathbf{S})}{\sqrt{\text{Var}[\mathbf{w}\mathbf{S}]}}}{\partial w_i^2} &= \frac{(\text{sign}(w_i)R(\mathbf{S}_i)2w_i - \sum_{l=1}^K |w_l|R(\mathbf{S}_l) - |w_i|R(\mathbf{S}_i)) \|\mathbf{w}\|^3}{\|\mathbf{w}\|^6} \\
&\quad - \frac{(\text{sign}(w_i)R(\mathbf{S}_i)\|\mathbf{w}\|^2 - w_i \sum_{l=1}^K R(\mathbf{S}_l)) 3w_i \|\mathbf{w}\|}{\|\mathbf{w}\|^6} \\
&= \frac{|w_i|R(\mathbf{S}_i) - R(\mathbf{S}_i)\|\mathbf{w}\|^2 / |w_i|}{\|\mathbf{w}\|^3} \\
&= \frac{R(\mathbf{S}_i)}{|w_i| \|\mathbf{w}\|^3} (w_i^2 - \|\mathbf{w}\|^2). \tag{3.101}
\end{aligned}$$

□

3.8.10 Proof of Lemma 19 (wording p. 151)

To compute the partial derivatives of $\tilde{\mathcal{C}}_R$ given by Eq. (2.42), we note that

$$d|W_{ij}|/dW_{ij} = \text{sign}(W_{ij}) \quad \text{if } W_{ij} \neq 0,$$

and that from [Petersen and Pedersen, 2005]

$$\frac{\partial \log |\det \mathbf{W}|}{\partial W_{ij}} = \left[\frac{\partial \log |\det \mathbf{W}|}{\partial \mathbf{W}} \right]_{ij} = \left[\mathbf{W}^{-1T} \right]_{ij} = W^{ij}.$$

Let us compute the partial derivative of W^{ij} with respect to W_{kl} . We note that $W^{ij} = \text{Tr}(\mathbf{E}_{ij} \mathbf{W}^{-1})$ where \mathbf{E}_{ij} is the matrix with only one nonzero element at

the (i, j) place which equals 1 and Tr denotes the trace, hence [Petersen and Pedersen, 2005]

$$\begin{aligned} \frac{\partial^2 \log |\det \mathbf{W}|}{\partial W_{kl} \partial W_{ij}} &= \frac{\partial \text{Tr}(\mathbf{E}_{ij} \mathbf{W}^{-1})}{\partial W_{kl}} \\ &= -[(\mathbf{W}^{-1} \mathbf{E}_{ij} \mathbf{W}^{-1})^T]_{kl} = -[\mathbf{W}^{-1} \mathbf{E}_{ij} \mathbf{W}^{-1}]_{lk} = -W^{il} W^{kj}. \end{aligned}$$

This yields the formula for the partial second order derivatives of the restriction of $\tilde{\mathcal{C}}_R$ to $\mathcal{M}_I(K)$ as given by the lemma.

3.8.11 Proof of Corollary 15 (wording p. 152)

By Lemma 20, for the restriction of $\tilde{\mathcal{C}}_R$ to $\mathcal{M}_I(K)$ to admit a local maximum point, it is necessary that the sections I_1, \dots, I_K of I be all disjoint. On the other hand, none of these sections can be all empty since otherwise $\mathcal{M}_I(K)$ would be empty. Therefore these sections must be reduced to a single point: $I_i = \{(i, j_i)\}$, $i = 1, \dots, K$ where j_1, \dots, j_K are indexes in $\{1, \dots, K\}$. These indexes must be distinct since otherwise $\mathcal{M}_I(K)$ would be empty. But a matrix in $\mathcal{M}_I(K)$ where $I = \{(1, j_1), \dots, (i, j_K)\}$ with j_1, \dots, j_K being a permutation of $1, \dots, K$, is simply a product of a diagonal and a permutation matrix and is thus similar to \mathbf{I}_K . Such a matrix is already known to realize the global maximum of $\tilde{\mathcal{C}}_R$. This completes the proof.

3.8.12 Proof of Lemma 20 (wording p. 152)

Let $\mathbf{W} \in \mathcal{M}_I(K)$ be a stationary point (if it exists) of the restriction of $\tilde{\mathcal{C}}_R$ to $\mathcal{M}_I(K)$, we shall show that it cannot realize a local maximum of this function. By assumption, there exists i, j, k in $\{1, \dots, K\}$ and $i \neq j$, such that (i, k) and (j, k) are both in I . Therefore, by Corollary 15, W^{ik} and W^{jk} are non-zero, hence by Lemma 19: $\partial^2 \tilde{\mathcal{C}}_R / \partial W_{ik} \partial W_{jk} = -W^{ik} W^{jk} \neq 0$. Also, by the same corollary, $\partial^2 \tilde{\mathcal{C}}_R / (\partial W_{ik})^2 = \partial^2 \tilde{\mathcal{C}}_R / (\partial W_{jk})^2 = 0$. Thus, let $\tilde{\mathbf{W}}$ be a matrix differing (slightly) from \mathbf{W} only at the indexes (i, k) and (j, k) : $\tilde{W}_{ik} = W_{ik} + \epsilon$, $\tilde{W}_{jk} = W_{jk} + \eta$, then since the first order partial derivatives of $\tilde{\mathcal{C}}_R$ vanish at \mathbf{W} , a second order Taylor expansion yields:

$$\tilde{\mathcal{C}}_R(\tilde{\mathbf{W}}) = \tilde{\mathcal{C}}_R(\mathbf{W}) - W^{ik} W^{jk} \epsilon \eta + O((|\epsilon| + |\eta|)^3)$$

as $\epsilon, \eta \rightarrow 0$. Therefore, $\tilde{\mathcal{C}}_R(\tilde{\mathbf{W}}) > \tilde{\mathcal{C}}_R(\mathbf{W})$ if ϵ and η both are small enough and their product have the same sign as $W^{ik} W^{jk}$. This shows that \mathbf{W} cannot realize a local maximum of $\tilde{\mathcal{C}}_R$ in $\mathcal{M}_I(K)$.

3.8.13 Proof of Lemma 21 (wording p. 152)

It is obvious that $\mathbf{W}\mathbf{W}^T \in \mathbb{R}^{P \times P}$ is symmetric. It is also invertible since it is full-rank : $\text{rank}(\mathbf{W}\mathbf{W}^T) = \text{rank}(\mathbf{W}) = P$.

Computation of first order derivatives

$$\frac{\partial \log |\det(\mathbf{WW}^T)|}{\partial W_{ij}} = \frac{1}{|\det(\mathbf{WW}^T)|} \frac{\partial |\det(\mathbf{WW}^T)|}{\partial W_{ij}} \quad (3.102)$$

But, noting the *trace* operator by $\text{Tr}(.)$:

$$\frac{\partial |\det(\mathbf{WW}^T)|}{\partial W_{ij}} = \det(\mathbf{WW}^T) \text{Tr} \left((\mathbf{WW}^T)^{-1} \frac{\partial (\mathbf{WW}^T)}{\partial W_{ij}} \right) . \quad (3.103)$$

Further, note that

$$\frac{\partial (\mathbf{WW}^T)}{\partial W_{ij}} = \mathbf{W} \frac{\partial \mathbf{W}^T}{\partial W_{ij}} + \frac{\partial \mathbf{W}}{\partial W_{ij}} \mathbf{W}^T \quad (3.104)$$

yielding

$$\frac{\partial (\mathbf{WW}^T)}{\partial W_{ij}} = \mathbf{W} \mathbf{J}^{ji} + \mathbf{G}^{ij} \mathbf{W}^T . \quad (3.105)$$

In the above equality, $\mathbf{J}^{ji} \in \mathbb{Z}^{K \times P}$ and $\mathbf{G}^{ij} \in \mathbb{Z}^{P \times K}$ are *single-entry* matrices : $[\mathbf{J}^{ji}]_{kl} = [\mathbf{G}^{ij}]_{lk} = \delta_{kj} \delta_{li}$, $(k, l) \in \mathbb{Z}^{K \times P}$ (only the (i, j) -th entry of both \mathbf{J}^{ji} and \mathbf{G}^{ij} matrices is non-zero, and is set to one).

Observe that for any matrices \mathbf{U}, \mathbf{V} with appropriated size:

$$\text{Tr}(\mathbf{V} \mathbf{J}^{ji}) = [\mathbf{V}^T]_{ij} \quad (3.106)$$

and

$$\text{Tr}(\mathbf{V} \mathbf{G}^{ij} \mathbf{U}) = [\mathbf{U} \mathbf{V}]_{ji} \quad (3.107)$$

Then, one gets

$$\begin{aligned} \text{Tr} \left((\mathbf{WW}^T)^{-1} \frac{\partial (\mathbf{WW}^T)}{\partial W_{ij}} \right) &= \text{Tr} ((\mathbf{WW}^T)^{-1} \mathbf{W} \mathbf{J}^{ji}) \\ &\quad + \text{Tr} ((\mathbf{WW}^T)^{-1} \mathbf{G}^{ij} \mathbf{W}^T) \\ &= \underbrace{\left[((\mathbf{WW}^T)^{-1} \mathbf{W})^T \right]}_{\doteq \mathbf{W}^+}_{ji} + \underbrace{\left[\mathbf{W}^T (\mathbf{WW}^T)^{-1} \right]}_{\doteq \mathbf{W}^+}_{ji} \\ &= 2[\mathbf{W}^+]_{ji} = 2[(\mathbf{W}^+)^T]_{ij} \end{aligned} \quad (3.108)$$

□

Computation of second order derivatives

$$\begin{aligned} \frac{\partial^2 \log |\det(\mathbf{WW}^T)|}{\partial W_{kl} \partial W_{ij}} &= \frac{\partial}{\partial W_{kl}} \text{Tr} \left((\mathbf{WW}^T)^{-1} \frac{\partial (\mathbf{WW}^T)}{\partial W_{ij}} \right) \\ &= \text{Tr} \left(\frac{\partial}{\partial W_{kl}} \left[(\mathbf{WW}^T)^{-1} \frac{\partial (\mathbf{WW}^T)}{\partial W_{ij}} \right] \right) . \end{aligned}$$

But, from (3.105)

$$\begin{aligned}
\text{Tr} \left(\frac{\partial}{\partial W_{kl}} \left[(\mathbf{W}\mathbf{W}^T)^{-1} \frac{\partial(\mathbf{W}\mathbf{W}^T)}{\partial W_{ij}} \right] \right) &= \text{Tr} \left(\frac{\partial [(\mathbf{W}\mathbf{W}^T)^{-1} (\mathbf{W}\mathbf{J}^{ji} + \mathbf{G}^{ij}\mathbf{W}^T)]}{\partial W_{kl}} \right) \\
&= \text{Tr} \left(\frac{\partial [(\mathbf{W}\mathbf{W}^T)^{-1} \mathbf{W}\mathbf{J}^{ji}]}{\partial W_{kl}} \right) \\
&\quad + \text{Tr} \left(\frac{\partial \left[(\mathbf{W}\mathbf{W}^T)^{-1} \underbrace{\mathbf{G}^{ij}\mathbf{W}^T}_{(\mathbf{W}\mathbf{J}^{ji})^T} \right]}{\partial W_{kl}} \right) \\
&= \text{Tr} \left(\frac{\partial [(\mathbf{W}\mathbf{W}^T)^{-1}]}{\partial W_{kl}} \mathbf{W}\mathbf{J}^{ji} \right) \\
&\quad + \text{Tr} \left(\frac{\partial [(\mathbf{W}\mathbf{W}^T)^{-1}]}{\partial W_{kl}} (\mathbf{W}\mathbf{J}^{ji})^T \right) \\
&\quad + \text{Tr} \left((\mathbf{W}\mathbf{W}^T)^{-1} \frac{\partial [\mathbf{W}\mathbf{J}^{ji}]}{\partial W_{kl}} \right) \\
&\quad + \text{Tr} \left((\mathbf{W}\mathbf{W}^T)^{-1} \frac{\partial [(\mathbf{W}\mathbf{J}^{ji})^T]}{\partial W_{kl}} \right) .
\end{aligned}$$

But, denoting \mathbf{H}^{ki} the single-entry matrix in $\mathbb{R}^{P \times P}$:

$$\text{Tr} \left((\mathbf{W}\mathbf{W}^T)^{-1} \frac{\partial [\mathbf{W}\mathbf{J}^{ji}]}{\partial W_{kl}} \right) = \delta_{lj} \text{Tr}[(\mathbf{W}\mathbf{W}^T)^{-1} \mathbf{H}^{ki}] \quad (3.109)$$

$$= \delta_{lj} [(\mathbf{W}\mathbf{W}^T)^{-1}]_{ki} , \quad (3.110)$$

and similarly, $\text{Tr} \left((\mathbf{W}\mathbf{W}^T)^{-1} \frac{\partial [(\mathbf{W}\mathbf{J}^{ji})^T]}{\partial W_{kl}} \right) = \delta_{lj} [(\mathbf{W}\mathbf{W}^T)^{-1}]_{ik}$. Consequently,

$$\text{Tr} \left((\mathbf{W}\mathbf{W}^T)^{-1} \frac{\partial [(\mathbf{W}\mathbf{J}^{ji})^T]}{\partial W_{kl}} \right) + \text{Tr} \left((\mathbf{W}\mathbf{W}^T)^{-1} \frac{\partial [\mathbf{W}\mathbf{J}^{ji}]}{\partial W_{kl}} \right) = 2\delta_{lj} [(\mathbf{W}\mathbf{W}^T)^{-1}]_{ik} .$$

But, noting that

$$\frac{\partial \mathbf{F}^{-1}}{\partial W_{ij}} = -\mathbf{F}^{-1} \frac{\partial \mathbf{F}}{\partial W_{ij}} \mathbf{F}^{-1} \quad (3.111)$$

we can see that

$$\text{Tr} \left(\frac{\partial [(\mathbf{W}\mathbf{W}^T)^{-1}]}{\partial W_{kl}} \mathbf{W}\mathbf{J}^{ji} \right) + \text{Tr} \left(\frac{\partial [(\mathbf{W}\mathbf{W}^T)^{-1}]}{\partial W_{kl}} (\mathbf{W}\mathbf{J}^{ji})^T \right)$$

equals

$$-\left\{ \text{Tr} \left[(\mathbf{WW}^T)^{-1} \frac{\partial(\mathbf{WW}^T)}{\partial W_{kl}} (\mathbf{WW}^T)^{-1} \mathbf{WJ}^{ji} \right] + \text{Tr} \left[(\mathbf{WW}^T)^{-1} \frac{\partial(\mathbf{WW}^T)}{\partial W_{kl}} (\mathbf{WW}^T)^{-1} (\mathbf{WJ}^{ji})^T \right] \right\} .$$

that is from Eq. (3.105):

$$-\text{Tr} [(\mathbf{WW}^T)^{-1} [\mathbf{WJ}^{lk} + \mathbf{G}^{kl} \mathbf{W}^T] (\mathbf{WW}^T)^{-1} [\mathbf{WJ}^{ji} + \mathbf{G}^{ij} \mathbf{W}^T]] .$$

The last expression is equal to the following sum of four traces :

$$\text{Tr} \left[\underbrace{(\mathbf{WW}^T)^{-1} \mathbf{W}}_{(\mathbf{W}^+)^T} \mathbf{J}^{lk} \underbrace{(\mathbf{WW}^T)^{-1} \mathbf{W}}_{(\mathbf{W}^+)^T} \mathbf{J}^{ji} \right] + \quad (3.112)$$

$$\text{Tr} \left[\underbrace{(\mathbf{WW}^T)^{-1} \mathbf{W}}_{(\mathbf{W}^+)^T} \mathbf{J}^{lk} (\mathbf{WW}^T)^{-1} \mathbf{G}^{ij} \mathbf{W}^T \right] + \quad (3.113)$$

$$\text{Tr} \left[(\mathbf{WW}^T)^{-1} \mathbf{G}^{kl} \underbrace{\mathbf{W}^T}_{\mathbf{W}^+} (\mathbf{WW}^T)^{-1} \mathbf{WJ}^{ji} \right] + \quad (3.114)$$

$$\text{Tr} \left[(\mathbf{WW}^T)^{-1} \mathbf{G}^{kl} \underbrace{\mathbf{W}^T}_{\mathbf{W}^+} (\mathbf{WW}^T)^{-1} \mathbf{G}^{ij} \mathbf{W}^T \right] . \quad (3.115)$$

To deal with the above traces, observe that

$$\text{Tr}[\mathbf{AJ}^{pq}\mathbf{BJ}^{rs}] = [\mathbf{A}]_{sp}[\mathbf{B}]_{qr} \quad (3.116)$$

and

$$\text{Tr}[\mathbf{AJ}^{pq}\mathbf{BJ}^{rs}\mathbf{C}] = \sum_{m=1} [\mathbf{A}]_{mp}[\mathbf{B}]_{qr}[\mathbf{C}]_{sm} \quad (3.117)$$

This yields the following equalities:

$$\begin{aligned} \text{Tr} [(\mathbf{W}^+)^T \mathbf{J}^{lk} (\mathbf{W}^+)^T \mathbf{J}^{ji}] &= [\mathbf{W}^+]_{li} [\mathbf{W}^+]_{jk} \\ \text{Tr} [(\mathbf{WW}^T)^{-1} \mathbf{G}^{kl} \mathbf{W}^+ \mathbf{WJ}^{ji}] &= [(\mathbf{WW}^T)^{-1}]_{ik} [\mathbf{W}^+ \mathbf{W}]_{lj} \\ \text{Tr} [(\mathbf{W}^+)^T \mathbf{J}^{lk} (\mathbf{WW}^T)^{-1} \mathbf{G}^{ij} \mathbf{W}^T] &= \sum_m [\mathbf{W}^+]_{lm} [(\mathbf{WW}^T)^{-1}]_{ki} [\mathbf{W}]_{mj} \\ &= [\mathbf{W}^+ \mathbf{W}]_{lj} [(\mathbf{WW}^T)^{-1}]_{ki} \\ \text{Tr} [(\mathbf{WW}^T)^{-1} \mathbf{G}^{kl} \mathbf{W}^+ \mathbf{G}^{ij} \mathbf{W}^T] &= \sum_m [(\mathbf{WW}^T)^{-1}]_{mk} [\mathbf{W}^+]_{li} [\mathbf{W}]_{mj} \\ &= [\mathbf{W}^+]_{li} [\mathbf{W}^+]_{jk} . \end{aligned}$$

Finally, since the inverse of a symmetric matrix is symmetric, $[(\mathbf{WW}^T)^{-1}]_{ki} = [(\mathbf{WW}^T)^{-1}]_{ik}$, and it comes that

$$\frac{\partial^2 \log |\det(\mathbf{WW}^T)|}{\partial W_{kl} \partial W_{ij}} = 2 \left\{ [(\mathbf{WW}^T)^{-1}]_{ki} (\delta_{jl} - [\mathbf{W}^+ \mathbf{W}]_{lj}) - [\mathbf{W}^+]_{li} [\mathbf{W}^+]_{jk} \right\} .$$

□

3.8.14 Proof of Lemma 23 (wording p. 154)

Let $\mathbf{W} \in \mathcal{M}_I^{P \times K}$ be a stationary point (if it exists) of the restriction of $\tilde{\mathcal{C}}_R$ to $\mathcal{M}_I^{P \times K}$. We shall show that it cannot realize a local maximum of this function.

Consider the case where $\cup_{i=1}^P I_i$ contains more than P elements. Then there must exist an index $j \in \cup_{i=1}^P I_i$ for which \mathbf{e}_j , the j -th row of the identity matrix of order K , is not contained in the linear subspace spanned by the rows of \mathbf{W} , since this subspace of dimension P . By definition, there exists $i \in \{1, \dots, P\}$ such that $(i, j) \in I$. Let $\tilde{\mathbf{W}}$ be a matrix differing from \mathbf{W} only in the entry W_{ij} by ϵ . Then by the Taylor expansion up to second order, noting that the first partial derivative of $\tilde{\mathcal{C}}_R$ vanishes at \mathbf{W} and using Corollary 17,

$$\tilde{\mathcal{C}}_R(\tilde{\mathbf{W}}) = \tilde{\mathcal{C}}_R(\mathbf{W}) + \frac{1}{2} [(\mathbf{WW}^T)^{-1}]_{ii} (1 - [\mathbf{W}^+ \mathbf{W}]_{jj}) \epsilon^2 + O(|\epsilon|^3) , \quad (3.118)$$

as $\epsilon \rightarrow 0$. It can be checked that $\mathbf{W}^+ \mathbf{W}$ is idempotent (i.e. $(\mathbf{W}^+ \mathbf{W})^2 = \mathbf{W}^+ \mathbf{W}$) and symmetric, and hence the same is true for $\mathbf{I}_K - \mathbf{W}^+ \mathbf{W}$. Thus the j -th diagonal element of $\mathbf{I}_K - \mathbf{W}^+ \mathbf{W}$, which is $1 - [\mathbf{W}^+ \mathbf{W}]_{jj}$, is the same as the squared norm of its j -th row. Therefore, $1 - [\mathbf{W}^+ \mathbf{W}]_{jj} \geq 0$ with equality if and only if the j -th row of $\mathbf{I}_K - \mathbf{W}^+ \mathbf{W}$ vanishes, or equivalently $\mathbf{e}_j = \mathbf{e}_j \mathbf{W}^+ \mathbf{W}$. But since \mathbf{e}_j is not in the linear subspace spanned by the rows of \mathbf{W} , this cannot happen. On the other hand, \mathbf{WW}^T is symmetric and positive definite, implying that so is its inverse, and thus there exists a nonsingular matrix \mathbf{P} such that $\mathbf{PP}^T = (\mathbf{WW}^T)^{-1}$. Consequently, each (i, i) entry of $(\mathbf{WW}^T)^{-1}$, which is the squared norm of the i -th row of \mathbf{P} , is strictly positive. Hence, the second term of the right hand side of (3.118) is strictly positive, yielding $\tilde{\mathcal{C}}_R(\tilde{\mathbf{W}}) > \tilde{\mathcal{C}}_R(\mathbf{W})$ for all $\epsilon \neq 0$ and small enough; \mathbf{W} is not a local maximum of $\tilde{\mathcal{C}}_R$ on $\mathcal{M}_I^{P \times K}$.

Consider now the case $I_i \cap I_j \neq \emptyset$ for some $i \neq j$ in $\{1, \dots, P\}$. Let $k \in I_i \cap I_j$. By Lemma 22: $\partial^2 \tilde{\mathcal{C}}_R / (\partial W_{ik} \partial W_{jk}) = [(\mathbf{WW}^T)^{-1}]_{ji} (1 - [\mathbf{W}^+ \mathbf{W}]_{kk}) - [\mathbf{W}^+]_{ki} [\mathbf{W}^+]_{kj}$. Also, by Corollary 17, $\partial^2 \tilde{\mathcal{C}}_R / (\partial W_{ik})^2 = [(\mathbf{WW}^T)^{-1}]_{ii} (1 - [\mathbf{W}^+ \mathbf{W}]_{kk})$ and $\partial^2 \tilde{\mathcal{C}}_R / (\partial W_{jk})^2 = [(\mathbf{WW}^T)^{-1}]_{jj} (1 - [\mathbf{W}^+ \mathbf{W}]_{kk})$. Thus, let $\tilde{\mathbf{W}}$ be a matrix differing (slightly) from \mathbf{W} only at the indexes (i, k) and (j, k) : $\tilde{W}_{ik} = W_{ik} + \epsilon$, $\tilde{W}_{jk} = W_{jk} + \eta$, then since the first order partial derivatives of $\tilde{\mathcal{C}}_R$ vanish at \mathbf{W} , a second order Taylor expansion yields:

$$\begin{aligned} \tilde{\mathcal{C}}_R(\tilde{\mathbf{W}}) &= \tilde{\mathcal{C}}_R(\mathbf{W}) \\ &\quad + \frac{1 - [\mathbf{W}^+ \mathbf{W}]_{kk}}{2} [\epsilon \ \eta] \begin{bmatrix} [(\mathbf{WW}^T)^{-1}]_{ii} & [(\mathbf{WW}^T)^{-1}]_{ij} \\ [(\mathbf{WW}^T)^{-1}]_{ji} & [(\mathbf{WW}^T)^{-1}]_{jj} \end{bmatrix} \begin{bmatrix} \epsilon \\ \eta \end{bmatrix} \\ &\quad - \epsilon \eta [\mathbf{W}^+]_{ki} [\mathbf{W}^+]_{kj} + O((|\epsilon| + |\eta|)^3) \end{aligned} \quad (3.119)$$

as $\epsilon, \eta \rightarrow 0$.

We have shown that $1 - [\mathbf{W}^+ \mathbf{W}]_{kk} \geq 0$. Further, from the positive definiteness of $(\mathbf{W} \mathbf{W}^T)^{-1}$, one gets

$$\begin{bmatrix} \epsilon & \eta \end{bmatrix} \begin{bmatrix} [(\mathbf{W} \mathbf{W}^T)^{-1}]_{ii} & [(\mathbf{W} \mathbf{W}^T)^{-1}]_{ij} \\ [(\mathbf{W} \mathbf{W}^T)^{-1}]_{ji} & [(\mathbf{W} \mathbf{W}^T)^{-1}]_{jj} \end{bmatrix} \begin{bmatrix} \epsilon \\ \eta \end{bmatrix} \geq 0 .$$

Therefore $\tilde{\mathcal{C}}_R(\tilde{\mathbf{W}}) \geq \tilde{\mathcal{C}}_R(\mathbf{W}) - \epsilon\eta[\mathbf{W}^+]_{ki}[\mathbf{W}^+]_{kj} + O((|\epsilon| + |\eta|)^3)$ implying that $\tilde{\mathcal{C}}_R(\tilde{\mathbf{W}}) > \tilde{\mathcal{C}}_R(\mathbf{W})$ for $\epsilon\eta$ having opposite sign to $[\mathbf{W}^+]_{ki}[\mathbf{W}^+]_{kj}$ and $|\epsilon| + |\eta| > 0$ and small enough. This proves that \mathbf{W} cannot realize a local maximum of $\tilde{\mathcal{C}}_R$.

□

3.8.15 Proof of Corollary 18 (wording p. 155)

By Lemma 23, for the restriction of $\tilde{\mathcal{C}}_R$ to $\mathcal{M}_I^{P \times K}$ to admit a local maximum point, it is necessary that the sections I_1, \dots, I_P of I be all disjoint and their union have at most P elements. On the other hand, none of these sections can be empty since otherwise $\mathcal{M}_I^{P \times K}$ would be empty. Therefore these sections must reduce to a single point: $I_i = \{(i, j(i))\}$, $i = 1, \dots, P$ where $j(1), \dots, j(P)$ are distinct indexes in $\{1, \dots, K\}$. By definition $j(i)$ denotes the column index of the unique non-zero elements of the i -th row of \mathbf{W} . Thus \mathbf{W} has a single non zero element per row and at most one nonzero element per column, meaning that $\mathbf{W} \in \mathcal{W}^{P \times K}$. Hence, a necessary condition for \mathbf{W} to be a local maximum point of $\mathcal{C}_R(\mathbf{B})$ is that $\mathbf{B}\mathbf{A} \in \mathcal{W}^{P \times K}$. This concludes the proof since, from Theorem 16, it is also a sufficient condition.

□

CHAPTER 4

MINIMUM RANGE AND LEAST ABSOLUTE BOUND METHODS

Abstract. Many general-purpose ICA algorithms offer an appealing trade-off between performance and speed. However, when some a priori knowledge is available on the sources, one could decide to rather exploit a specific contrast, *fitted* to the properties of the sources. For instance, objective functions exploiting sparsity, non-negativity or finite measure support of the sources have been proposed in the scientific literature; other constraints can also be dealt with. Any a priori information regarding source signals may help to improve BSS algorithms, from at least four different viewpoints:

- To improve speed and convergence rate.
- To improve separation in terms of a performance index.
- To relax assumption (such as the source independence or on the square size of \mathbf{A}).
- To derive cost functions with more interesting properties (e.g. discriminant).

For example, if the sources are bounded, the minimum support-based algorithm may satisfy, depending of the context, the last three points. From all the information-theoretic contrasts that have been analyzed in the previous chapters, that is the contrasts based on Rényi's entropies, only the extended Rényi entropy with $r = 0$ is a discriminant contrast. Therefore, a detailed analysis of this criterion seems interesting.

Contribution. We propose geometrical interpretations of the range-based contrast function and discusses its practical use. The use of averaged quasi-ranges is suggested and a result of Chu giving a bound on the difference of order-statistics is used to propose a way for choosing a default value for the parameter of the averaged quasi-range range estimator. A simple optimization algorithm is proposed for the maximization of non-differentiable contrasts. We show why the criterion successes in separating correlated images if some post-processing is applied in the case of two images. The optimization over the orthogonal group is then extended to avoid cumulation errors in deflation: the rigid constraint is replaced by a smooth constraint. Finally, the criterion is extended to sources that are bounded on one side only. Part of this work (sections 4.3.1, 4.4.1.2 and 4.4.2) results from a close collaboration with my friend and former colleague J.A. Lee.

Part of the results presented in this chapter was or will be published in the following papers (see Appendix B): JA2, ICB3, ICB4, ICB5, ICP9, ICP11, ICP12.

Organization of the chapter. In this chapter, the geometrical interpretation of the criterion is first sketched for the simultaneous and deflation extraction schemes. Next, some statistical methods for estimating the extreme points of a distribution are reminded, from which the range can be deduced. Yet another one, suitable for our purposes is proposed; a default value for the single parameter and simple minimization algorithms are also provided. The impact of the source independence assumption is analyzed through a simple example involving correlated images separation. Finally, a non-orthogonal extension of the separation method, as well as the extension to only upper- or lower-bounded signals reveals to be promising, even on large-scale and low sample data sets, with possible dependency between the sources.

4.1 GEOMETRIC INTERPRETATION OF THE MINIMUM RANGE APPROACH

Looking at the scatter plots of two independent and bounded sources, of two non-trivial linear mixtures of them, or of the whitening of these mixtures provides an intuitive view of how the bounded sources can be recovered. Because the sources are bounded and independent, the boundary of the scatter plot is clearly a rectangle with edges parallel to the source axes and the corners being located at $(\inf S_1, \inf S_2)$, $(\inf S_1, \sup S_2)$, $(\sup S_1, \inf S_2)$, $(\sup S_1, \sup S_2)$. Obviously, even if the edges of the joint pdf form a rectangle, the pdf itself is not necessarily uniform in the rectangle; it depends on the marginal sources densities as $p_{S_1, S_2}(x, y) = p_{S_1}(x)p_{S_2}(y)$ (as an example, one source could have a sinusoidal temporal structure). This is shown in Fig 4.1.(a) for two white uniform sources. Mixing these signals through a mixing matrix yields a parallelepiped with edges parallel to the columns of \mathbf{A} (Fig. 4.1.(b)). Consequently BSS aims here at find-

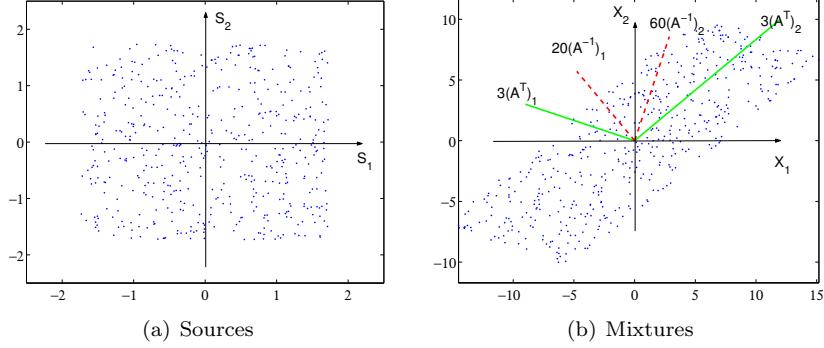


Figure 4.1. Scatter plots of independent sources and their linear mixtures.

ing a transformation \mathbf{B} such that the edges of the joint density $p_{\mathbf{B}_1 \mathbf{x}, \mathbf{B}_2 \mathbf{x}}(u, v)$ form a rectangle aligned with the axes. As usual, this can be done either by using a deflation or a simultaneous approach.

4.1.1 Interpretation of the simultaneous approach

From the above considerations, the direction of the columns of the mixing matrix \mathbf{A} can be recovered by estimating the edges of the convex hull of the mixture pdf, which is a parallelepiped. This method was proposed in [Prieto et al., 1998]. The major problem with this approach is its relatively high computational complexity.

We adopt here another point of view, leading to the simpler criterion $\mathcal{C}_R(\mathbf{B})$, which is proved to have interesting geometrical interpretations. We propose here two of them, adopting different viewpoints.

4.1.1.1 Interpretation in the mixture space

In the output space, the surface of the smallest rectangle (dashed) including the joint pdf of the outputs (gray) having its edges parallel to the orthogonal axes is given by $\prod_{i=1}^K R(\mathbf{Y}_i)$ (see Fig. 4.2.(a)). In the mixture space, as $\mathbf{X} = \mathbf{B}^{-1}\mathbf{Y}$, this rectangle looks like a parallelepiped of volume $V_{\mathbf{X}}(\mathbf{B}) = \prod_{i=1}^K R(\mathbf{Y}_i)/|\det \mathbf{B}|$; this is illustrated (dashed lines) in Fig. 4.2.(c) This parallelepiped contains the joint pdf of the mixtures (gray parallelepiped). It is amusing (and useful) to note that $\mathcal{C}_R(\mathbf{B}) = -\log V_{\mathbf{X}}(\mathbf{B})$. Hence, by maximizing $\mathcal{C}_R(\mathbf{B})$, we are looking for a matrix \mathbf{B}^* that yields the smallest dashed parallelepiped including the gray one (which represents the mixture pdf); this solution is shown in Fig. 4.2.(d) The point is that the matrix \mathbf{B}^* is PD-equivalent to \mathbf{A}^{-1} ; \mathbf{B}^* maps the mixture pdf to an output pdf whose convex hull is a rectangle with edges parallel to the output axes; in the output space, this solution looks like Fig. 4.2.(b) The remaining indeterminacies correspond to the usual \mathbf{PD} matrix : the order of

the outputs does not change the volume of the “mixture” parallelepiped, which further depends on the product $|\det \mathbf{A}| \prod_{i=1}^K R(\mathbf{S}_i)$ only.

4.1.1.2 Interpretation in the output space

In the general case where K independent and bounded sources are involved, the volume $V_Y(\mathbf{B})$ of the gray parallelepiped of Fig. 4.2.(a) (i.e. the volume of the convex hull of the output pdf) is equal to

$$V_Y(\mathbf{B}) = |\det(\mathbf{AB})| \prod_{i=1}^K R(\mathbf{S}_i) , \quad (4.1)$$

while the volume of the hyper-rectangle including this parallelepiped equals $\prod_{i=1}^K R(\mathbf{Y}_i)$ (dashed rectangles in Fig. 4.2.(a)). The point is that it can be shown that if the gray parallelepiped has a fixed volume (which corresponds to the $|\det \mathbf{B}| = cst$ constraint), the dashed rectangle will have a lowest volume (given by the product of the orthogonally projected lengths of the gray parallelepiped onto an orthogonal basis, which are nothing else than the corresponding ranges) when the gray parallelogram has a rectangular shape, with edges parallel to the basis axes (see Fig. 4.2.(b)). We are precisely looking for a separating matrix \mathbf{B}^* that corresponds to the last transformation subject to $|\det \mathbf{B}| = c'$ or equivalently $|\det \mathbf{W}| = c$ where c, c' are some constants. These constants do not matter since $|\det \mathbf{W}| = |\det \mathbf{B} \det \mathbf{A}|$ and $|\det \mathbf{A}|$ cannot be estimated (only the volume of $|\det \mathbf{A}| \prod_{i=1}^K R(\mathbf{S}_i)$ is known): the indeterminacy about the value of $|\det \mathbf{W}|$ results from the indeterminacy on the source ranges.

In our S-BSS $\mathcal{C}_R(\mathbf{B})$ criterion defined in Eq. (2.41), there is no constraint about the value of $|\det \mathbf{B}|$ (no “subject to” restriction to the search space¹), so the above explanations only correspond to a specific separation scheme in which $|\det \mathbf{B}|$ is kept constant. In that case indeed, we are looking for a demixing matrix with a given determinant that maps the gray parallelepiped of Fig. 4.2.(c) to a scaled copy of the gray rectangle in Fig. 4.2.(b) (the scaling depends on the fixed value of the determinant). In a more general way, what we are actually doing when maximizing $\mathcal{C}_R(\mathbf{B})$ is nothing else than minimizing a weighted sum of $\log \prod_{i=1}^K R(\mathbf{Y}_i)$ and $\log |\det \mathbf{B}|$; the last term prevents the singularity of the separating matrix (the criterion equals ∞ in this case) that would correspond to a “zero-volume shape” (a point or a line). But can we interpret the additive nature of the constraint? Why this specific form? Why equal weights? The matrices being stationary points of $\tilde{\mathcal{C}}_R(\mathbf{W})$ may have an arbitrary determinant: multiplying any of its row has no impact on its stationary point specificity; it still fulfills the stationary point condition, and this is quite natural as $\det \mathbf{A}$ cannot be recovered. Consequently, what we would like to do is not exactly to fix the value of the determinant. We can see $\log |\det \mathbf{B}|$ (resp. $\log |\det \mathbf{W}|$) as

¹even if one could restrict the search space to $\mathcal{SO}(K)$ if a prewhitening step is performed, as already explained

a penalization term in $\mathcal{C}_R(\mathbf{B})$ (resp. $\tilde{\mathcal{C}}_R(\mathbf{W})$), and $\tilde{\mathcal{C}}_R(\mathbf{W})$ can be thought of as the following penalized objective function:

$$\sum_{i=1}^K \log \left[\sum_{j=1}^K |W_{ij}| R(S_j) \right] - \lambda \log |\det \mathbf{W}| . \quad (4.2)$$

Any transformation of matrix \mathbf{W} can be seen as changing the direction of the rows as well as their scale. And what we would like to achieve actually is to compensate exactly the variation of $\sum_{i=1}^K \log \left[\sum_{j=1}^K |W_{ij}| R(S_j) \right]$ to the row scaling. The point is that $\log |\det \mathbf{W}|$ and $\sum_{i=1}^K \log \left[\sum_{j=1}^K |W_{ij}| R(S_j) \right]$ vary exactly at a same rate to a change of scaling; both $|\det \mathbf{W}|$ and $R(\mathbf{w}_i S)$ are linear in the norm of \mathbf{w}_i . By contrast, the penalization term is invariant under rotation. This explains why $\lambda = 1$: the coefficient does not fix the value of $\det \mathbf{W}$, but it fixes the rate of the change such that $\lambda \log |\det \mathbf{W}|$ compensates the effects of the norm of the rows of \mathbf{W} on $\sum_{i=1}^K \log \left[\sum_{j=1}^K |W_{ij}| R(S_j) \right]$.

Lets us come back to geometry. The above reasoning means that the volume of the dashed rectangle can only be minimized by changing the directions of the edges of the light gray parallelograms. In particular, if the direction of the edges are modified by a rotation transform, the volume of the light gray parallelepiped will again be kept constant because $V_Y(\mathbf{B})$ remains unchanged (see Eq. (4.1)), but the product of the output ranges can change since the penalization term vanishes. On the contrary, the penalization term compensates exactly the effect of stretching the rows of \mathbf{W} on the volume of the dashed rectangle, but not exactly the volume variation due to a modification of the directions of the rows of \mathbf{W} .

It is of interest to note that the linear transformation that maximizes $\mathcal{C}_R(\mathbf{B})$ are the non-singular matrices having each of their rows perpendicular to $K - 1$ columns of \mathbf{A} (see Fig. 4.1.(b)). The method proposed by Prieto et al. consisted in i) estimating the edges of the scatter plots and then ii) computing their perpendicular. This is rather heavy. Our analysis shows that actually, it suffices to minimize $\prod_{i=1}^K R(Y_i)$ without changing the volume of the gray parallelepiped, that is to minimize $V_X(\mathbf{B})$, which is a scalar quantity.

4.1.2 Interpretation of the deflation approach

Further geometrical aspects of the criterion can be emphasized when focusing on the deflation approach.

Let us suppose we have a centered rectangle and let us analyze the width of the projection of this rectangle onto a rotating vector as a function of the rotation angle ω . One shall face a local minimum when and only when the vector is orthogonal to one of the edges of the rectangle, and a local maximum when and only when the vector points to a corner; this directly results from the Pythagorean theorem. Clearly, the global minimum is obtained when \mathbf{w} is orthogonal to the edges of the rectangle with minimum inter-distances.

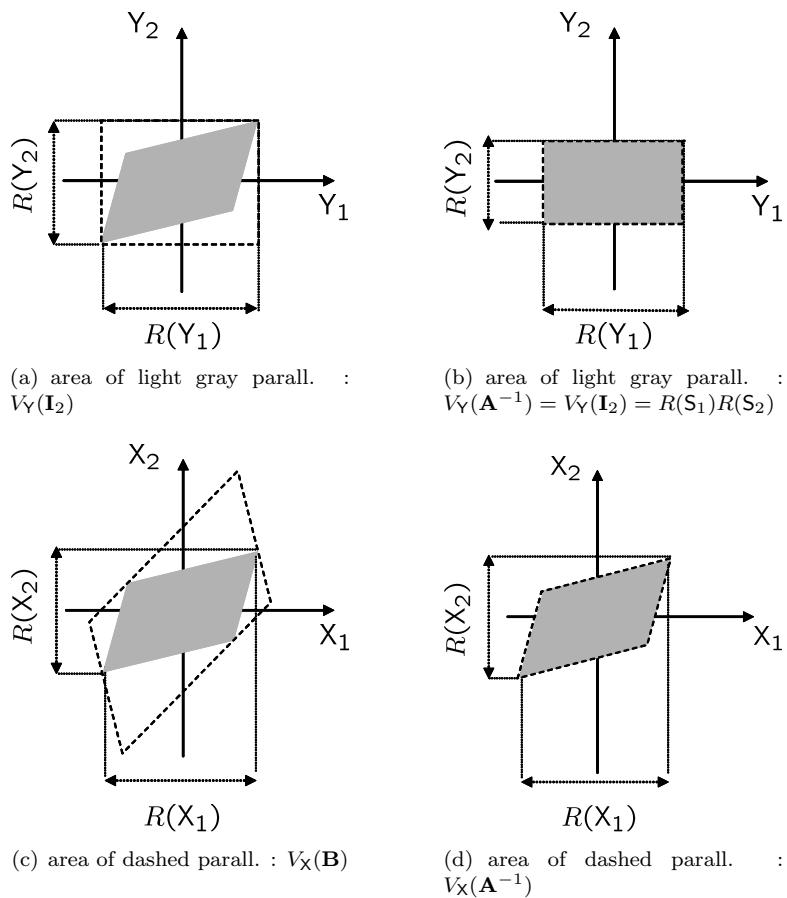


Figure 4.2. Geometric interpretation of $\mathcal{C}_R(\mathbf{B})$ in the output (top) and mixture (bottom) spaces ($\mathbf{A} \approx \mathbf{I}_2$).

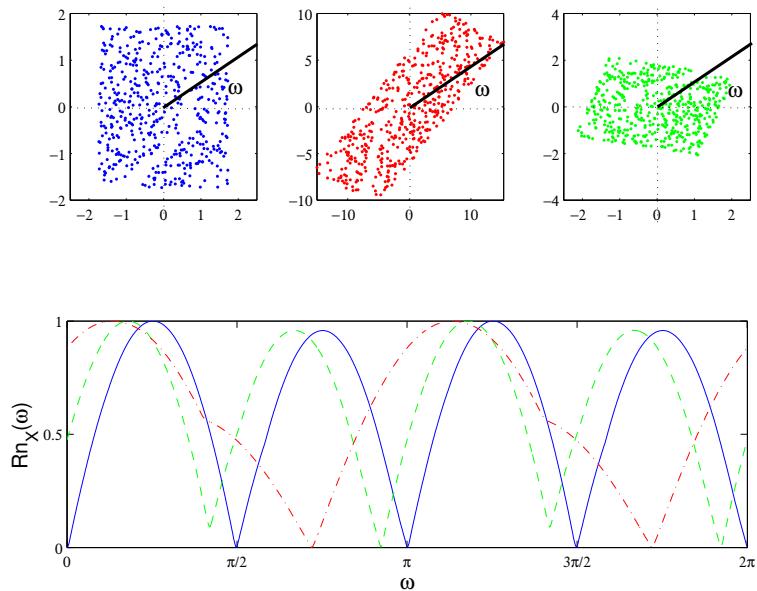


Figure 4.3. top:scatter plot of the sources ($\mathbf{A} = \mathbf{A}_1$, left), mixtures ($\mathbf{A} = \mathbf{A}_2$, middle) and whitened mixtures ($\mathbf{A} = \mathbf{A}_3$, right); bottom: evolution vs ω of $R_{nx}(\omega)$ with $\mathbf{A} = \mathbf{A}_1$ (solid), $\mathbf{A} = \mathbf{A}_2$ (dash-dot) and $\mathbf{A} = \mathbf{A}_3$ (dashed).

Actually, when one is modifying a row of the demixing matrix, we observe the same phenomenon when we look at the corresponding output range, which corresponds to the projection width of the rectangle onto the vector. This is shown in Figure 4.3. for the pair of uniform sample sources vector shown in Fig. 4.1.(a) and the three mixing matrices $\mathbf{A}_1 = \mathbf{I}_K$, $\mathbf{A}_2 = [-3, 6; 1, 5]$ and \mathbf{A}_3 a rotation matrix of angle close to $-\pi/15$. The top plots show the mixtures resulting from the various mixing matrices (from left to right). The bottom graph shows the normalized range (i.e. normalized projection widths of the parallelepipeds). More precisely, if we denote by $R_\omega(\mathbf{X})$ the range of the output $[\cos \omega, \sin \omega]\mathbf{X}$ for short, the angular plot can be normalized to $[0, 1]$ by applying the transform

$$Rn_{\mathbf{X}}(\omega) \doteq \frac{R_{\mathbf{X}}(\omega) - \min_{\omega}(R_{\mathbf{X}}(\omega))}{\max_{\omega}(R_{\mathbf{X}}(\omega) - \min_{\omega}(R_{\mathbf{X}}(\omega)))} . \quad (4.3)$$

The solid, dash-dotted and dashed curves at the bottom of Fig. 4.1.(a) respectively correspond to $\mathbf{X} = \mathbf{A}_1\mathbf{S}$, $\mathbf{X} = \mathbf{A}_2\mathbf{S}$ and $\mathbf{X} = \mathbf{A}_3\mathbf{S}$. It is shown that $Rn_{\mathbf{A}_1\mathbf{S}}(\omega)$ reaches its local minimum value when the edges of the square are parallel to the axes and its local maximum point when the corners are aligned with the axis, as expected from the geometrical insights. Further, if the sources have a symmetric pdf with ranges $R(\mathbf{S}_1) = 2a$ and $R(\mathbf{S}_2) = 2b$, the edges of the pdf form a rectangle centered at the origin and the projection widths evolve as $\sqrt{a^2 + b^2}/\cos(\omega - \arctan(b/a))$ for $\omega \in [0, \pi/2]$ and as $\sqrt{a^2 + b^2}/\cos(\pi - \omega - \arctan(b/a))$ in $[\pi/2, \pi]$ (there is a symmetry of order π); all the local maxima should equal the length of the corresponding diagonal of the rectangle. The quantity $Rn_{\mathbf{A}_3\mathbf{S}}(\omega)$ follows a similar law as $Rn_{\mathbf{A}_1\mathbf{S}}(\omega)$ because the scatter plot of $\mathbf{A}_3\mathbf{S}$ is a rotated hyper-rectangle; the remaining rotation, i.e. the shift between the dashed and solid curves in the bottom plot of Fig. 4.3. is due to the mixing angle caused by the mixing-whitening steps². What happens with $\mathbf{A} = \mathbf{A}_2$ in Fig. 4.3. seems more strange. Indeed, one could think that a local maximum should be observed when ω points to a corner and a local minimum when it is perpendicular to any column of \mathbf{A} . This is not always the case, as proved by adopting again a geometrical viewpoint.

Consider the parallelepipeds shown in Fig. 4.4.; they all have the same volume hb (here, we have set $h = b = 1$). The “projection widths”, noted $L(\omega)$, of the parallelepiped onto a vector with direction $\omega \in [-\pi/2, \pi/2]$ for various values of the “squeeze angle” $0 < \alpha \leq \pi/2$ are shown in Fig. 4.5.(b); the value of b and h have been kept constant so that the surface is unchanged as α varies (see Fig. 4.4.). Depending on the squeeze angle, the number of local maximum varies. When $\alpha < \pi/4$ (i.e. $\alpha < \arctan(h/b)$) only the “outer corners” (those joined by the largest diagonal of the parallelepiped, noted D in what follows) induce a local maximum, while the corners being closer from the origin (located

²Observe that contrarily to the theory, the dashed and solid curves are not exactly shifted copies of each other. Normally, the local minima must all be equal as they correspond to the range of a unit-variance signal with uniform density (up to the normalization); this results from a unperfect whitening based on samples only.

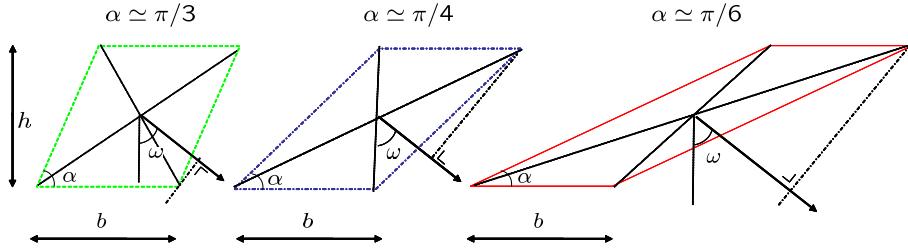


Figure 4.4. Parallellepipeds for various squeeze angles α with constant volume bh .

on the shortest diagonal, noted d in the following) do no more induce a local maximum. Similarly, the local minimum points corresponding to ω orthogonal to the shortest sides (i.e. $\omega \in \{k\pi | k \in \mathbb{Z}\}$) also vanish in this situation; this is emphasized in Fig. 4.5.(a) where the maximum projection widths are shown for $\alpha \in \{\pi/6, \pi/4, \pi/3, \pi/2\}$ (here: $h = b = 1$). The theoretical $L(\omega, \alpha)$ projections are shown. By noting $b' = L - b$ where L is the width of the parallelogram projected onto the horizontal line (that is, the projection width of the parallelepiped onto the rotating vector of angle $\omega = \pi/2$: $L \doteq L(\pi/2, \alpha)$), these lengths are given by the following formulas (each term in the max is the projection of a diagonal of the parallelogram onto the rotating vector ω):

$$\left\{ \begin{array}{l} \max \left(\sqrt{1 + (1+b')^2} \cdot |\cos(\pi/2 - \omega + \arctan(\frac{1}{1+b'}))|, \right. \\ \quad \left. \sqrt{1 + (1-b')^2} \cdot |(\cos(\arctan(\frac{1}{1-b'})) - \omega)| \right), \text{ if } \alpha \geq \pi/4, \\ \\ \max \left(\sqrt{1 + (1+b')^2} \cdot |\cos(\pi/2 - \omega + \arctan(\frac{1}{1+b'}))|, \right. \\ \quad \left. \sqrt{1 + (1-b')^2} \cdot |\cos(\arctan(\frac{1}{1-b'}) + \omega)| \right), \text{ if } \alpha \leq \pi/4. \end{array} \right.$$

Indeed, depending on the mixing matrix, only the maximum points corresponding to the corners belonging to D are observed, and only the local minimum points corresponding to ω perpendicular to the pair of parallel edges (the largest sides) being very close to each other are preserved (see Fig. 4.5.(b)).

Because of the relationship between i) the parallelepiped and the mixture density convex hull in one hand and ii) the projection widths and the output ranges on the other hand, these results may seem to be contradictory to Theorem 15 given in p. 63, which basically states that there is a local minimum point of the output range under a fixed variance constraint. This is not the case when things are appropriately compared. Let us focus on the simple $K = 2$ case for the ease of the illustration. Computing $\tilde{\mathcal{C}}_R(\mathbf{w}_\theta)$ w.r.t. θ reduces to computing the projection widths of a rectangle (parallelogram with $\alpha = \pi/2$) onto a unit-norm rotating vector \mathbf{w} . This is exactly the same as computing $\mathcal{C}_R(\mathbf{b}_\theta)$ w.r.t. θ onto the unit-norm rotating vector $\mathbf{b}_\theta \doteq [\cos \theta, \sin \theta]$ when the mixture pdf convex hull is a rotated rectangle (that is under prewhitening). What happens when the mixtures are not whitened? When no prewhitening is performed, the

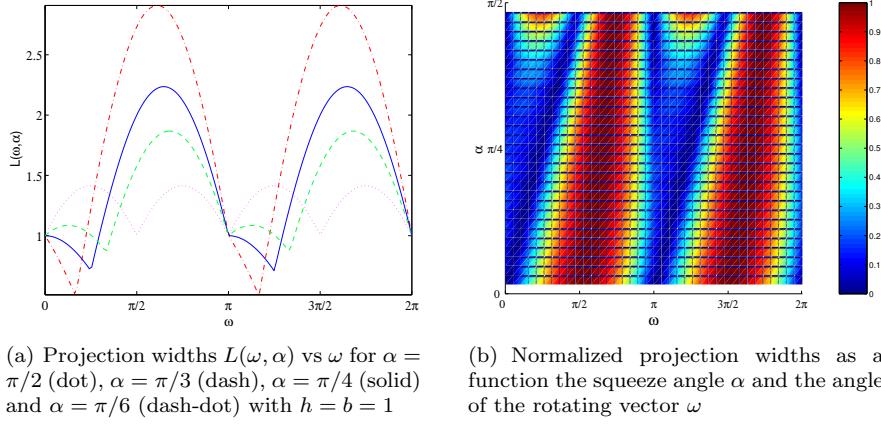


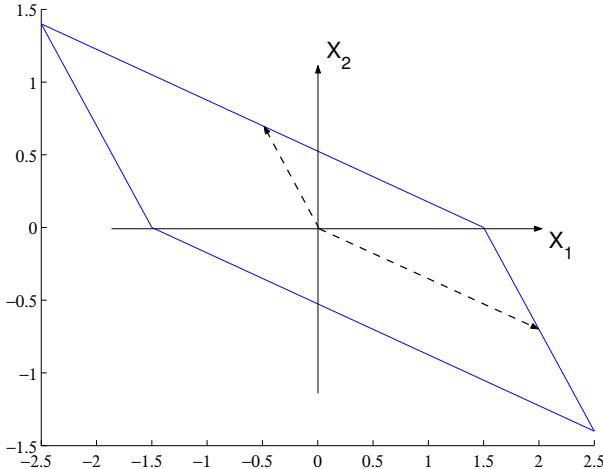
Figure 4.5. Effect of the squeeze angle α on the number of local minima of the range.

normalization constraint becomes $\text{Var}[\mathbf{b}\mathbf{X}] = \mathbf{b}\Sigma_{\mathbf{X}}\mathbf{b}^T = 1$. The point is that we are no longer searching for projections onto a fixed-length rotating vector ! The vector \mathbf{b} is stretched with respect to $\Sigma_{\mathbf{X}}$ and consequently, the rotating vector must have a varying length in order that $\text{Var}[\mathbf{b}\mathbf{X}] = cst$. This is illustrated in the following example.

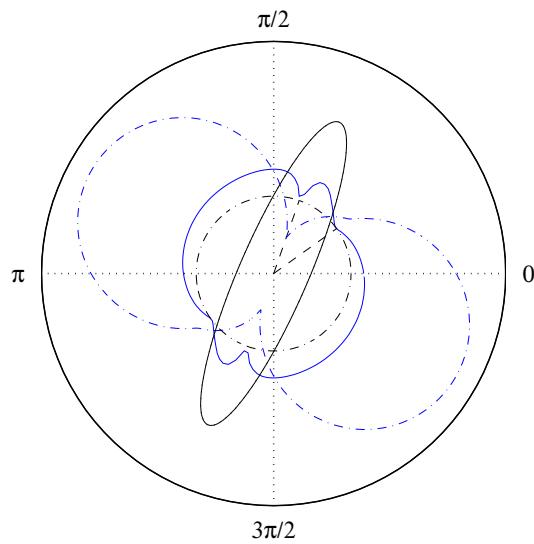
Example 26 (Local minimum point and projection widths) Consider the mixing matrix $\mathbf{A} = [2, -0.5; -0.7, 0.7]$; under a unitary source covariance matrix Σ_S , the mixture covariance matrix is given by $\Sigma_{\mathbf{X}} = \mathbf{A}\mathbf{A}^T$. The parallelogram shown in Fig. 4.6.(a) shows the mixture pdf convex hull where the (fictive) source ranges were $[-1, 1]$ (in other words, when the above mixing matrix is applied to the unit-square, one obtains this parallelogram). The arrows indicate the direction of the columns of \mathbf{A} . Figure 4.6.(b) shows the evolution of $R(\mathbf{b}\mathbf{X})$. Depending on the constraint on \mathbf{b} , the corresponding manifolds are either the dash-dotted circle ($\mathbf{b}\mathbf{b}^T = 1$) or the solid ellipsoid ($\mathbf{b}\Sigma_{\mathbf{X}}\mathbf{b}^T = 1$). The straight lines (“radius-like”) plotted in the first quadrant are given by the perpendicular directions to the edges of the parallelogram of Fig. 4.6.(a); each is seen to point to a local minimum point of $R(\mathbf{b}\mathbf{X})$ if, indeed, $\mathbf{b}\Sigma_{\mathbf{X}}\mathbf{b}^T = 1$. This is not the case if $\|\mathbf{b}\| = 1$ on this example, because the “squeeze angle” α of the parallelogram is too small.

The above example shows that we recover the local minima once a pair of edges of the parallelogram is perpendicular to the axes provided that the rotating vector \mathbf{b}_{θ} is scaled by $1/\sqrt{\mathbf{b}_{\theta}\Sigma_{\mathbf{X}}\mathbf{b}_{\theta}^T}$, which depends on θ if $\Sigma_{\mathbf{X}} \neq \mathbf{I}_K$.

In summary, the geometrical interpretation of the deflation procedure is rather intuitive. Let us focus on the simple 2D case for visualization purposes. The edges of the source scatter plot form a rectangle. The effect of the non-singular mixing matrix is to stretch the “source” rectangle into a “mixture” parallelepiped. When extracting a first source, one is looking for the rotating



(a) Mixture pdf convex hull and the direction given by the columns of the mixing matrix.



(b) Sets $S(2)$ (dash-dot circle) and $\mathbf{b}\Sigma_X\mathbf{b}^T = 1$ (solid ellipsoid). The projected widths of the parallelogram on the corresponding rotating vectors \mathbf{b} are shown in dash-dotted and solid, respectively.

Figure 4.6. Example 26: effect of the constraint of \mathbf{b} on the non-mixing local minimum points. One can see on the right panel that a local minimum of the width of the projected parallelogram onto the rotating vector \mathbf{b} is observed in any direction perpendicular to the columns of the mixing matrix (shown by the arrows on the left panel) if the $\mathbf{b}\Sigma_X\mathbf{b}^T = 1$ constraint is used, some minima disappear if $\|\mathbf{b}\| = 1$ is used instead.

vector \mathbf{b}_1 such that the projection of the parallelepiped onto the vector \mathbf{b}_1 has a minimum width (that is, the signal $\mathbf{b}_1 \mathbf{X}$ has a minimum range). As this vector satisfies $\mathbf{b}_1 \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{b}_1^T = 1$, its norm must vary with the rotation angle. Such a vector can be written as $\mathbf{b}_1 = \mathbf{b}_\theta / \sqrt{\mathbf{b}_\theta \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{b}_\theta^T}$.

The above vectors \mathbf{b}_1 corresponding to the directions perpendicular to the edges of the hyper-parallelepiped correspond to a local minimum of the range of the projected signals.

4.2 RANGE ESTIMATION

This section definitely does not aim at presenting a detailed review of range estimation techniques. Neither theoretical nor empirical comparison between these methods will be performed. The reason is that our final goal is not range estimation, but well to achieve BSS through the maximization of a range-based contrast function. Therefore, we shall only remind some of the most known range estimation techniques (or endpoint estimation techniques from which a range estimator can be found) and emphasize their limitations in our BSS context. Next a new estimator will be proposed, in which the value of the single parameter can be appropriately chosen without knowing the density of the parent random variable. The proposed estimator is simple, but its simplicity allows us to point out the problems that could be encountered in our BSS context when range estimation is poorly managed. In spite of that, this estimation technique has led us to promising results in terms of separation performances (analyzing more efficient -and probably more complicated- estimators could be the purpose of a further work). Relationship with the above existing estimators will also be investigated. In the remaining of the chapter, we focus on the range estimation of a random variable X , having well-defined continuous density and distribution functions, with support convex hull of the form $\bar{\Omega}(X) = (a, b)$ with $-\infty < a < b < \infty$.

Range estimation can be based on extreme points estimation (also called endpoint estimation), which is a well-known problem in statistics and econometrics. Many estimators have been proposed in this framework, and their asymptotic behavior have been analyzed. Some of them are recalled in the next subsection.

4.2.1 Some existing methods for endpoint estimation

In this section, we briefly recall some methods for estimating the endpoint θ of a distribution function P_R satisfying $P_R(0) = 0$ and $P_R(\theta) = 1$; we further assume that for all $\epsilon > 0$, $P_R(\epsilon) > 0$ and $P_R(\theta - \epsilon) < 1$ so that this distribution function could be the cdf of a random variable R with support convex hull $\bar{\Omega}(R) = [0, \theta]$. Assume that a sample set \mathcal{R} of size N is available: $\mathcal{R} = \{r_1, \dots, r_N\}$ where each r_i is an observed sample of the random variable R ; one can build an ordered set \mathcal{R}' from these measurements $\mathcal{R}' = \{r_{(1:N)}, \dots, r_{(N:N)}\}$ where $r_{(i:N)} \leq r_{(i+1:N)}$

are the ordered elements of \mathcal{R} . In the statistical literature, the above $r_{(i:N)}$ are called the *i-th order statistic of \mathcal{R}* (i.e. its *i-th* largest element). This is actually a slight abuse of notation since the “order statistic” expression also refers to a random variable rather than to a sample point. More precisely, the “*i-th order statistic*” appellation is also used for the random variable $R_{(i:N)}$, which represents the rv describing the value of the *i-th* largest element in a sample set \mathcal{R} of size N , each of the component of \mathcal{R} being drawn from the parent density $p_R(r)$ [Feller, 1966]. This is because when analyzing the theoretical behavior of a functional involving $r_{(i:N)}$, the variable $R_{(i:N)}$ are used instead.

Most probably, the simplest way to estimate the endpoint θ is to approximate it by its largest observed value in the sample set, that is $\theta \approx \theta_N \doteq r_{(N:N)}$. In this case however, only a single point is used in the estimation. Furthermore, in the noise-free case, $\theta_N < \theta$ for finite N because the probability to observe $R = \theta$ is zero since R is a continuous random variable. Therefore, it seems appropriate to artificially “enlarge” this value: $\theta \approx \theta_N + \rho(N)$. Intuitively speaking, the enlargement $\rho(N)$ should depend on the last *spacings*, that is on the differences between successive points near the boundary of the density. But when N increases, the samples tend to “fill the support of the density”, and $\rho(N)$ should decrease monotonously with N . Quenouille proposed in 1949 the following expression for $\rho(N)$, which corresponds to a estimator of θ [Quenouille, 1949]:

$$\rho(N) = \frac{N-1}{N} (r_{(N:N)} - r_{(N-1:N)}) \quad (\text{Quenouille 49}) , \quad (4.4)$$

it relies thus on the two largest sample points. This seems to be suboptimal in the sense that more robust methods could be obtained by using more than two points. Moreover, these extreme points can be outliers. In 1979, Peter Cooke suggested a radically opposite method [Cooke, 1979]. It is still of the form $\theta \approx \theta_N + \rho(N)$, but $\rho(N)$ involves now each of the sample points:

$$\rho(N) = \theta_N - \sum_{i=0}^{N-1} \left\{ (1 - i/N)^N - \left(1 - \frac{i+1}{N}\right)^N \right\} r_{(N-i:N)} \quad (\text{Cooke 79}) . \quad (4.5)$$

One year later the same author proposes in [Cooke, 1980] a compromise: an m -points method is investigated to approximate θ :

$$\hat{\theta}_{\text{Cooke-80}} \doteq \sum_{i=1}^m \alpha_i r_{(N-i+1:N)} \quad (\text{Cooke 80}) . \quad (4.6)$$

This estimator requires the computation of several parameters: obviously the integer m has to be fixed and the α_i coefficients have to be found. Cooke proposed to determine them in such a way that the above estimator has the smallest mean squared error (MSE), asymptotically as $N \rightarrow \infty$. This necessarily requires a model for the distribution $P_R(r)$ of R in a neighborhood of $r = \theta$, and an additional parameter (that he noted ν) had to be estimated, or at least fixed.

The major advantage of this method is that a closed-form solution for the vector $\alpha \doteq [\alpha_1, \dots, \alpha_m]$ can be obtained in terms of the Gamma function.

Two years later, in 1982, a maximum-likelihood estimator $\hat{\theta}_{\text{Hall}}$ [Hall, 1982] was proposed. The problem however is that the estimator is a solution of some equations, which analytical form depends on the behavior of the model density near θ . Despite the fact that the estimator is robust to a departure from the true range pdf to the model pdf used in the maximum likelihood approach, obtaining $\hat{\theta}_{\text{Hall}}$ is not an easy task, as it is the root of the following function

$$\frac{1}{\sum_{j=1}^{m-1} \log \left(1 + \frac{r_{(N-j+1:N)} - r_{(N-m+1:N)}}{\theta - r_{(N-j+1:N)}} \right)} - \frac{1}{\sum_{j=1}^{m-1} \frac{r_{(N-j+1:N)} - r_{(N-m+1:N)}}{\theta - r_{(N-j+1:N)}}} - \frac{1}{m} ,$$

with the additional condition that $\hat{\theta}_{\text{Hall}} > \theta_N$. Moreover this equation only admits at least one solution with probability one as $N \rightarrow \infty$.

Another approach has been proposed more recently in [Meister, 2006]. The endpoint θ is estimated by using the moments of R , based on the fact that the sequence $\sqrt[j]{E[R^j]}$ converges increasingly to θ with j . Taking a small value for j means that both small and large samples influence the estimator: the variance decreases, contrarily to the bias. Taking a large value for j amounts to attaching much more importance to large observed values: only few points dominate. Therefore, we can see the j exponent as a regularization parameter that controls the relative importance of the samples in the estimator depending of their value. For finite j , all values influence the value of the estimate and for an infinite value of j , only the largest value does matter.

4.2.2 Existing Range estimation

In spite of the fact that the above endpoint estimation techniques are range estimation of specific random variables (because when $\bar{\Omega}(R) = [0, \theta]$ estimating the range is the same as estimating the endpoint θ), they can be extended to the range estimation of random variables with support convex hull of finite measure of the form (a, b) . The random variable R could be, in our context, the observed range R_N (defined below) of the random variable X based on the sample set $\mathcal{X} = \{x_1, \dots, x_N\}$ (the x_i are assumed to be i.i.d.), which is always positive and finite; this justifies the notation of the considered random variable by the letter “R” in the previous subsection. The observed range is defined as the largest difference between two sample points in \mathcal{X} , i.e.

$$R_N \doteq \max_{i,j} \{|x_i - x_j|, x_i, x_j \in \mathcal{X}\} , \quad (4.7)$$

or equivalently, using the order statistics notation (defining \mathcal{X}' in a similar way as \mathcal{R}'):

$$R_N = x_{(N:N)} - x_{(1:N)}, x_{(1:N)}, x_{(N:N)} \in \mathcal{X}' . \quad (4.8)$$

Like for R , the random variable associated to the i -th largest sample in a set \mathcal{X} of observations of X of size N is noted by a capital letter $X_{(i:N)}$. We assume

that the samples x_i are drawn from a random variable X with one-dimensional continuous pdf $p_X(x)$ with finite range; our aim is now to estimate the range of X , that is if $\bar{\Omega}(X) = [a, b]$ with $R(X) = b - a < \infty$. We are looking for an estimate of $b - a$ from endpoint estimation techniques. Clearly, $R_N \in [0, \theta]$ with $\theta = b - a$ is the value that has to be estimated.

More generally, relationships between endpoint estimation of a positive random variable and range estimation of whatever bounded variable arises: the $r_{(i:N')}$ can be obtained from the set \mathcal{X} via difference between its elements, that is $r_{(i:N')} = |x_{(i':N)} - x_{(j':N)}|$ for some pair $(x_{(i':N)}, x_{(j':N)})$ in the sample set \mathcal{X} . Note that both \mathcal{R} and N' depend on the construction procedure of the $r_{(i:N')}$ from the $x_{(i:N)}$: for example, if all the non-trivial differences are considered, that is $\mathcal{R} = \{|x_{(i:N)} - x_{(j:N)}|, i \neq j\}$, then $N' = N(N - 1)/2$. But, if we build $\mathcal{R} = \{|x_{(N-i+1:N)} - x_{(i:N)}|, 1 \leq i \leq \lfloor N/2 \rfloor\}$, then $N' = \lfloor N/2 \rfloor$.

In the specific case of Moment-based Meister's approach, an estimation of $b - a$ is given by the j -th square root of the empirical mean of the set $\{|x_{p+\lfloor N/2 \rfloor} - x_p|^j, x_p \in \mathcal{X}, p \leq \lfloor N/2 \rfloor\}$. Meister's approach is thus approximatively based on $\lfloor N/2 \rfloor$ differences, and should be sensitive on how the sample set is splitted into two parts; he proposed a random choice, where the two subsets are built from the (meaningless) time indexes p of the samples x_p , but some more efficient choice could be also possible. For instance, why not defining \mathcal{R} as (a subset of) $\{\cup_{i=1}^{\lfloor N/2 \rfloor} (x_{(N-i+1:N)} - x_{(i:N)})\}$? Most probably, this is because the theoretical behavior of the estimator is more complicated to study: each of the spacings may have a different density. Indeed in the last differences, the indexes of the sample points are not arbitrary anymore. As an example, the probability to observe $x_{(N-i+1:N)} - x_{(i:N)}$ close to θ should be much higher for i close to one than for i close to $\lfloor N/2 \rfloor$ and, similarly, the probability to observe a value of $x_{(N-i+1:N)} - x_{(i:N)}$ close to 0 should be much higher for i close to $\lfloor N/2 \rfloor$ than for i close to one. On the contrary, as the samples are i.i.d., the time indexes of x_i are meaningless so that there is no reason for the distribution of $x_i - x_j$ to differ from that of $x_p - x_q$; there is no order relation between the samples.

Just as for endpoint estimation, we can derive from [Devroye and Wise, 1980] the following estimator of $b - a$, which is similar to endpoint estimators of the form $\theta = \theta_N + \rho(N)$:

$$R(X) = R_N + 2\rho(N) \quad (\text{Devroye \& Wise 80}) \quad , \quad (4.9)$$

where $\rho(N)$ is a functional parameter to be fixed. It should be decreasing in N ; large if the tails of the density p_X are smoothly decreasing and small if the probability to observe points near the boundaries is high (sharp tails).

In the last case, efficient approaches for estimating the frontiers (also called boundaries) of densities can be derived, possibly under measurement error; they rely on the essential assumption that support boundaries are jump discontinuity points of a density (for more details about this approach, we refer to [Hall and Simar, 2002, Delaigle and Gijbels, 2003] and references therein). In our application however, we would like not to rule out densities decaying (possibly

smoothly) at the support boundaries, such as e.g. “triangular” densities, even if this would mean that only less efficient range estimators can be found.

The main idea behind range estimation based on endpoint estimation techniques is to deal with a random variable describing differences between samples. There are other possible approaches however, as e.g. support estimation via pdf estimation, as briefly presented in the next subsection.

4.2.2.1 Support estimation via density estimation

One of the first ideas that probably comes in mind for estimating the range $R(X)$ of a rv X is to estimate its pdf $\hat{p}_X(x) \approx p_X(x)$ and then infer on its support by identifying the set where the approximated pdf “lives”, that is

$$R(X) \approx \mu[\overline{\{x : \hat{p}_X(x) > 0\}}] . \quad (4.10)$$

However, this is an ill-posed problem because pdf estimation is often managed using pointwise convergence concepts. For example, Parzen windowing is a universal approximator of continuous densities in the sense of L_p -norm or pointwise convergence under mild conditions on the window width σ_K (meaning basically that σ_K tends to zero not faster than $1/N$ as the number N of sample points tends to infinity). For instance we can say that \hat{p}_X is a good estimate of p_X in the sense that

$$D_A(p_X \| \hat{p}_X) = \int |p_X(x) - \hat{p}_X(x)| dx \quad (4.11)$$

tends to zero with probability one as $N \rightarrow \infty$ [Devroye and Györfi, 1985, Silverman, 1986]. But the problem is that the mapping between the pdf p_X to the support $\Omega(X)$ may be discontinuous. Just think about a random variable with support $\Omega(X) = [a, b]$ for finite $-\infty < a < b < \infty$. Estimating $\Omega(X)$ from a Parzen estimate \hat{p}_X of the pdf p_X is a non-sense if the support of the kernels used in Parzen estimation is e.g. the whole real line. Clearly, it is the case when, for example, the pdf is estimated via a weighted sum of Gaussian functions: the observed range derived from \hat{p}_X is infinite despite the fact that $D_A(p_X \| \hat{p}_X)$ can be made as small as desired. More generally, the range estimation is highly sensitive to the support of the kernel used in the pdf estimation. Of course, one could approximate

$$R(X) \approx \mu[\overline{\{x : \hat{p}_X(x) > \epsilon\}}] \quad (4.12)$$

with $\epsilon > 0$ but the value of ϵ might be difficult to guess and clearly, depends on the chosen kernel in case of kernel density estimation, as well as on the density p_X . In both cases, a new question arises “how to estimate the density in such a way that the associated (possibly approximated) range matches the original range”? This is a very difficult question actually and hence, we turn to other estimation techniques.

4.2.2.2 Range estimation for BSS application

The problem with most of the above range estimators, except Meister's moment-based technique, is that they are not really suitable for our BSS application: either they are too naïve (no parameters, relying on extremely few sample points and thus highly sensitive to noise) or some parameters depending on the density of the observed range (and thus indirectly on the parent source densities) have to be fixed. In BSS application, one does not want to spend a lot of computational time to estimate or to model source densities as efficient and fast methods exist to deal with the BSS problem. We want to obtain better separation performances but with similar computational complexity. Actually, the tricky tuning parameters result from i) the need to have some theoretical results about the estimator, and ii) the need to generalize to the much more difficult task of support estimation of possibly noisy random vectors (i.e. of rv taking values in higher-dimensional space). Some of them also involve computational-intensive resampling techniques (see e.g. [Loh, 1984]), which is also a major limitation in BSS.

Consequently, the above-described methods do not really match our requirements in BSS applications: they are either too simple or too complicated to tune and to compute; comparing them in great detail would be quite useless in the framework of this thesis. Therefore, we decided to use a dedicated estimator, which is very simple, intuitive and seems efficient for BSS applications, even if we admit that no theoretical and empirical comparison between the proposed and existing range estimators have been performed. However, connections between our estimator and some of the above estimation techniques can be pointed out.

4.2.3 Quasi-range based approach

In the remaining part of the thesis, we shall consider an order-statistics range estimator, based on the empirical mean of the set $\{x_{(N:N)} - x_{(1:N)}, \dots, x_{(N-m+1:N)} - x_{(m:N)}\}$, where $m < \lfloor N/2 \rfloor$ is the counterpart of the m parameter in $\theta_{\text{Cooke } 80}$. However, an advantage is that we focus on “relevant” differences: just like the information near the boundaries of the support is essentially needed to estimate the shape of pdf tails, only sample points near the boundaries (more exactly, the last spacings) should be of primary importance in the estimation of the boundary location. As it will be shown below, the proposed estimator has the great advantage that i) m has a very simple intuitive meaning (it is half the number of points considered in the estimator: the m smallest and m largest points of the sample set) and ii) a threshold value for m controlling the probability of making an error smaller than another threshold can be found without a priori knowledge on the density of the pdf p_x from which the samples have been drawn (see Section 4.2.3.3). This is especially convenient in blind applications like BSS.

We define the m -averaged order statistic differences by

$$\langle R_N^{<m>}(\mathbf{X}) \rangle \doteq \frac{1}{m} \sum_{p=1}^m R_N^{<p>}(\mathbf{X}), \text{ with } R_N^{<p>}(\mathbf{X}) \doteq x_{(N-p+1:N)} - x_{(p:N)}. \quad (4.13)$$

Obviously, $\langle R_N^{<m>}(\mathbf{X}) \rangle$ can also be seen as the difference between the averages of m -th first and last order statistics. The quantity $\langle R_N^{<m>}(\mathbf{X}) \rangle$ can be used as a simple estimator of the range $R(\mathbf{X})$. At first sight, from a statistician perspective, this estimator does not seem appealing: it is actually a lower bound of the true range for finite sample size. Nevertheless, its simplicity leads to a not time-consuming estimation technique, contrary to e.g. resampling methods. Furthermore, some connections can be emphasized with the aforementioned range estimators. With $m = 1$, $\langle R_N^{<1>}(\mathbf{X}) \rangle$ simply reduces to $R_N(\mathbf{X}) \doteq R_N^{<1>}(\mathbf{X})$, the observed range of the sample set $\{x_1, \dots, x_N\}$. It corresponds to the Devroye & Wise estimator (4.9) with $\rho(N) = 0$. More generally, it corresponds to $\hat{\theta}_{\text{Cooke-80}}$ if $\alpha_i = 1/m$ and the $r_{(i:N)}$ are deduced from the sample values of \mathbf{X} by $x_{(N-i+1:N)} - x_{(i:N)}$. Setting all the α_i to a same value of $1/m$ has advantages and drawbacks. The major drawback is that the estimator is a lower bound on the true range in absence of noise (it is even a lower bound on R_N). The advantage is that the estimation of an additional parameter that characterizes the shape of the tails of the density of $R_N(x)$ is avoided. In spite of the drawback, we shall focus in the following on the estimator $\langle R_N^{<m>}(\mathbf{X}) \rangle$ because, as it will be shown soon, it suits our needs in BSS applications. Estimators available from the statistical literature (see among other the work of Devroye, Cooke, Härdle, Simar, Loh, Hall, Tsybakov, etc) are good, but their final objective is not to be plugged in ICA methods. In the ICA application, a precise estimation of the support is not really needed. By contrast, the shape of this estimator versus the elements of the transfer matrix (i.e. when $K = 2$, the mixing angle), as well as the sensitivity to the size of the sample set and to measurement noise are the hot points; the ultimate goal is that the discriminacy and contrast properties still hold after the range estimation step. As an example, a possible “height shift” of the surface plotted in Fig. 3.26. has no impact on the set of local maximizers (i.e. on the obtained demixing matrix).

In the sequel, we shall first focus on $R_N(\mathbf{X})$ as a naïve estimate of $R(\mathbf{X})$ which is simple, fast and does not involve the tricky adjustment of some parameters. The analysis is then extended to $\langle R_N^{<m>}(\mathbf{X}) \rangle$ with $m > 1$ to improve robustness. As the order statistics have been extensively studied in the statistic literature, some interesting results can be given. In particular, a meaningful threshold value for m can be found based on existing statistical results.

4.2.3.1 The observed range estimator

When a theoretical study of the range estimator has to be performed the i -th largest value $x_{(i:N)}$ has to be replaced by the i -th order statistic $X_{(i:N)}$. When these random variables are used instead, the associated observed range becomes

a random variable as well. It will be noted as usual using the roman font (i.e. $R_N(X)$ instead of $R_N(X)$) and defined as:

$$R_N(X) \doteq X_{(N:N)} - X_{(1:N)} . \quad (4.14)$$

As for any random variable, we can compute its probability density function, which obviously depends on the “parent density”, that is of p_X . The theoretical density $p_{R_N(X)}$ is given by (see [Hoel, 1975, David, 1970] and in the Appendix of the chapter, Section 4.6.1)

$$p_{R_N(X)}(r) = N(N-1) \int_{\inf(X)}^{\sup(X)-r} p_X(u)p_X(u+r) \left[\int_u^{u+r} p_X(x)dx \right]^{N-2} du \quad (4.15)$$

Observe that if $\inf(X)$ and $\sup(X)$ are not isolated points, $\inf(X) = \lim_{N \rightarrow \infty} X_{(1:N)}$ and $\sup(X) = \lim_{N \rightarrow \infty} X_{(N:N)}$.

Having in mind Eq. (4.15), we can now compute the expectation and variance of the error $R(X) - R_N(X)$ ³. Both $E[R_N(X)]$ and $\text{Var}[R_N(X)]$ depend on p_X .

Example 27 Let U be a uniform r.v. in $(0, 1)$ with pdf $p_U(u) = 1$ for $u \in (0, 1)$. Then,

$$p_{R_N(U)}(r) = N(N-1)r^{N-2}(1-r) . \quad (4.16)$$

This result leads to

$$E[R_N(U)] = \frac{N-1}{N+1} , \quad (4.17)$$

and

$$\text{Var}[R_N(U)] = \frac{N(N-1)}{(N+1)(N+2)} - \frac{(N-1)^2}{(N+1)^2} \quad (4.18)$$

$$= \frac{2(N-1)}{(N+2)(N+1)^2} . \quad (4.19)$$

Let L be a r.v. following the linear law $p_L(x) = 2x$ with $x \in [0, 1]$. Then, tedious manipulations give

$$p_{R_N(L)}(r) = \frac{Nr^{N-1}}{2(N+1)} \left\{ \frac{(2-r)^{N-1}}{r} [(2-r)^2(N-1) - r^2(N+1)] + 2r^N \right\} . \quad (4.20)$$

³Absolute value symbols have been omitted since, in the noise-free case, we always have $\Pr(\inf(X) < x_i < \sup(X)) = 1$ for all $1 \leq i \leq N$ with $N < \infty$ implying that $0 < R_N(X) < R(X)$ with probability one.

In the linear pdf case, because of the difficulty to compute the analytical form of the first two moments of $R_N(L)$, a numerical integration of r and r^2 with respect to the above density $p_{R_N(L)}(r)$ will be preferred.

The performances of $R_N(X)$ as an estimator of $R(X)$ can be analyzed from various viewpoints. The bias and variance of $R_N(X)$ are computed. We also investigate the effects of the sample size N , the density p_X of X and of additive Gaussian noise $G_n \sim \mathcal{N}(0, \sigma_n^2)$.

In what follows, we adopt the following notation for the random variables. The symbols U, L, T and V represents random variables with uniform, linear, triangular and "V"-shape densities, respectively. These pdfs and the related cdfs are illustrated on Fig.4.7.

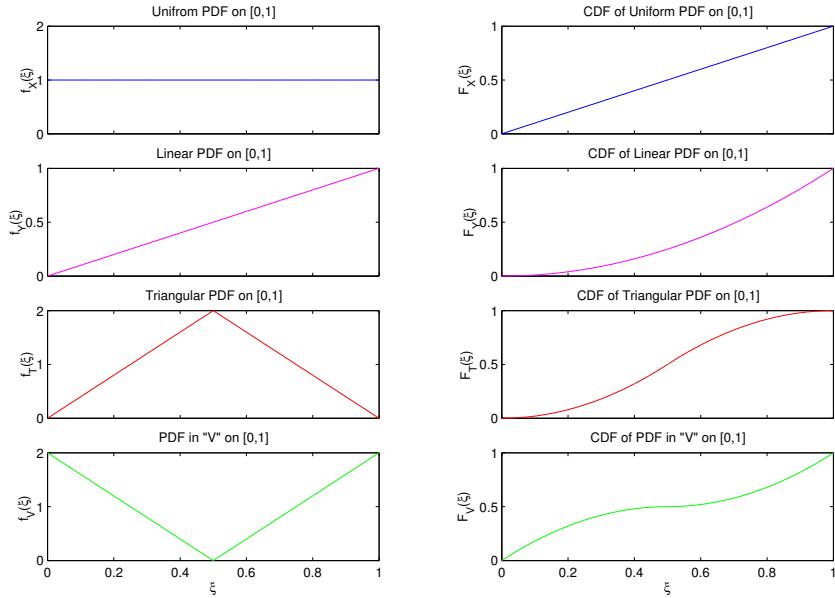


Figure 4.7. Densities and distributions of uniform (U), linear (L), triangular (T) and "V" (V) random variables defined on $(0, 1)$.

- Bias

Looking at Eq. (4.17), it is obvious that $R_N(U)$ is an asymptotically unbiased estimator of $R(U)$. For instance, if $\Omega(U) = (0, 1)$:

$$\lim_{N \rightarrow \infty} E[R_N(U)] = 1 . \quad (4.21)$$

Even if we are not able to extend this result to other densities (because of the difficulty to compute $E[R_N(X)]$ when p_X is not uniform), we conjecture

that if X is a random variable with support $\Omega(X)$ such that the support convex hull is of the form (a, b) , and if there exists $\epsilon > 0$, $\epsilon \in \mathbb{R}$ such that $(a + \epsilon) \in \Omega(X)$, $(b - \epsilon) \in \Omega(X)$, then $R_N(X)$ is an asymptotically unbiased estimator of $R(X)$; i.e. :

$$\lim_{N \rightarrow \infty} E[R_N(X)] = R(X) = b - a . \quad (4.22)$$

The above result shows that the mean value of the error

$$\mathcal{E}_N(X) \doteq R(X) - R_N(X) \quad (4.23)$$

tends to zero when the sample size increases.

- Consistency

Another estimator property is the consistency. The estimator $R_N(X)$ is consistent if it converges in probability to $R(X)$ with increasing N .

Definition 25 (Convergence in probability) A sequence X_1, X_2, \dots is said to converge in probability to a random variable X if for every $\epsilon > 0$

$$\lim_{N \rightarrow \infty} \Pr(|X_N - X| > \epsilon) = 0 .$$

It is rather intuitive that $\Pr[\mathcal{E}_N(X) < \epsilon] \leq \Pr[\mathcal{E}_{N'}(X) < \epsilon]$ for $N' > N$. The equality case only holds if $P_{R_{N'}(X)}(R(X) - \epsilon) = P_{R_N(X)}(R(X) - \epsilon)$, i.e. the cdf of $R_N(X)$ should not depend on N , which seems not natural. Therefore, it is reasonable to conjecture that for any $0 < \epsilon < R(X)$, if $N < N'$, then $\Pr[\mathcal{E}_N(X) < \epsilon] < \Pr[\mathcal{E}_{N'}(X) < \epsilon]$. Consequently, $\lim_{N \rightarrow \infty} \Pr[\mathcal{E}_N(X) < \epsilon] = 1$, which is the definition of the consistency of $R_N(X)$ as an estimate of $R(X)$.

Let us rigorously show that $R_N(U)$ is a consistent estimator of $R(U)$ (special case where $p_X = p_U$ is uniform). In this case:

$$\begin{aligned} \Pr[\mathcal{E}_N(U) < \epsilon] &= \Pr[R_N(U) - R(U) > -\epsilon] \\ &= 1 - P_{R_N(U)}(R(U) - \epsilon) \\ &= 1 - \frac{R(U) + \epsilon(N-1)}{R(U)} \left(\frac{R(U) - \epsilon}{R(U)} \right)^{N-1} . \end{aligned}$$

By the L'Hospital rule, the last chain of equalities leads to $\lim_{N \rightarrow \infty} \Pr[R(U) - R_N(U) < \epsilon] = 1$.

- Variance

The variance $\text{Var}[\mathcal{E}_N(X)]$ reduces to $\text{Var}[R_N(X)]$. When $p_X = p_U$, then

$$\lim_{N \rightarrow \infty} \text{Var}[R_N(X)] = 0 , \quad (4.24)$$

and again, we conjecture that this result holds for other densities under the same conditions as for the asymptotically unbiased property.

- Effect of $p_X(x)$

Even if e.g. linear random variables are not often encountered in real-world applications, the four random variables whose pdfs and cdfs are shown in Figure 4.7. permit us to emphasize the effect of the *density shape*.

In practice, the density (4.15) is difficult to use because of the difficulty to compute the integral. However, it is possible to deal with it in simple cases, as e.g. when X is a uniform or linear random variable.

Figure 4.8. illustrates that the density $p_{R_N(X)}(r)$ as a function of N clearly depends on the pdf p_X . We can see that the most probable values for $R_N(X)$ converge to one for both uniform and linear variable when N goes to infinity, but the rates of convergence are different. The above result is purely theoretical. Note that in the linear case, $E[R_N(X)]$ is estimated using a numerical integration of $p_{R_N(X)}$ with respect to r setting $p_X(x) = 2x$.

- Effect of the noise

The robustness to noise is expected to be very poor. Therefore we shall first consider another estimator that will be analyzed under various viewpoints, including robustness to additive Gaussian noise.

4.2.3.2 The m -averaged quasi-range estimator

In the above subsection, we have analyzed the performances of $\langle R_N^{<1>}(X) \rangle$ as an estimator of $R(X)$. In practice, this estimator can suffer from a lack of robustness in presence of additive noise, because it only relies on two “extreme” sample points. A simple way to modify this estimator is to take an average of order-statistic differences, i.e. to consider $\langle R_N^{<m>}(X) \rangle$ where m is an integer greater than or equal to 1 but smaller than $N/2 - 1$. As we have done for R_N , we will focus on basic estimator properties, in a theoretical way as far as possible⁴.

As for the empirical range, here are some comments about the performances of the estimator. The bias and the variance of the estimator are explicitly computed for the uniform parent density.

⁴Other estimators, than are not necessarily biased for finite N , like it was the case for $R_N^{<m>}$ or $\langle R_N^{<m>}(X) \rangle$ have been investigated in a similar way as Cooke has done, i.e. by adding a weighted sum of the last spacings. The corresponding results were disappointing in the sense that no improvement can be observed from the BSS separation results viewpoint, which is our target goal. On the contrary, the adjustment of additional parameters (i.e. the number of spacings and the corresponding weights) revealed to be tricky. They are not shown here to keep conciseness.

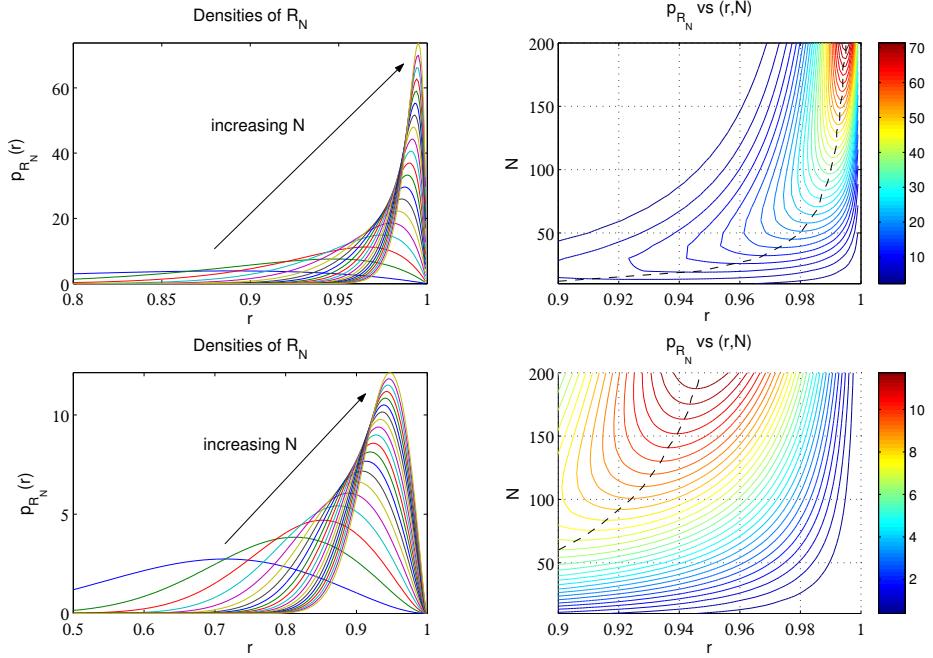


Figure 4.8. Evolution of $p_{R_N(\cdot)}(r)$ vs r (N is varying) for uniform (top left) and linear (bottom left) variables defined on $(0, 1)$. The joint pdf $p_{R_N(r)}(r, N)$ for uniform (top right) and linear (bottom right) r.v. are shown, and the most probable values of $p_{R_N(r)|N}$ are indicated by the dashed curve.

- Bias

As for the $m = 1$ case, $\langle R_N^{<m>}(X) \rangle$ with $1 < m \leq \lfloor N/2 \rfloor$ is an asymptotically unbiased estimator, when m does not depend on N . For example, if m is fixed then, using the Appendix of the chapter given in Section 4.6.2:

$$E[\langle R_N^{<m>}(U) \rangle] = \frac{1}{m} \sum_{p=1}^m \frac{N - 2p + 1}{N + 1} = \frac{N - m}{N + 1} . \quad (4.25)$$

By contrast, this is not necessarily true if m is a function of N , e.g. if m is an integer multiple of N . Assume $k = N/m \in \mathbb{Z}$:

$$\begin{aligned} \lim_{N \rightarrow \infty} E[\langle R_N^{<N/k>}(U) \rangle] &= \lim_{N \rightarrow \infty} k/N \left\{ \sum_{p=1}^{N/k} \frac{N - 2p + 1}{N + 1} \right\} \\ &= \lim_{N \rightarrow \infty} \left\{ 1 - \frac{2k/N}{N + 1} \sum_{p=1}^{N/k} p \right\} \\ &= 1 - \lim_{N \rightarrow \infty} \frac{N/k + 1}{N + 1} \\ &= 1 - 1/k . \end{aligned} \quad (4.26)$$

On the other hand, if m is of the form $(N/k)^q$ with $k > 0$, $0 < q < 1$, the estimator can be asymptotically unbiased.

We have not addressed the possible consistency property, due to the difficulty of computing $P_{\langle R_N^{<m>}(X) \rangle}$ for $m > 1$.

- Variance

We find easily that if we define

$$\mathcal{E}_N^{<m>}(.) \doteq R(.) - \langle R_N^{<m>}(.) \rangle , \quad (4.27)$$

then $\text{Var}[\mathcal{E}_N^{<1>}(U)] = \text{Var}[R_N(U)] = \frac{2(N-1)}{(N+2)(N+1)^2}$. More generally, $\text{Var}[\mathcal{E}_N^{<m>}(U)]$ can be rewritten as

$$\begin{aligned} \text{Var}[\langle R_N^{<m>}(U) \rangle] &= \frac{1}{m^2} \left\{ \sum_{p=1}^m \text{Var}[R_N^{<p>}(U)] \right. \\ &\quad \left. + 2 \sum_{1 \leq i < j \leq m} \text{Cov}[R_N^{<i>}(U), R_N^{<j>}(U)] \right\} . \end{aligned}$$

Noting that, from the Appendix given in Section 4.6.3, $\text{Cov}_{p < q} [U_{(p:N)}, U_{(q:N)}] = \frac{p(N+1-q)}{(N+2)(N+1)^2}$, we find:

$$\text{Cov}_{i < j} [R_N^{<i>}(U), R_N^{<j>}(U)] = 2i \frac{N+1-2j}{(N+2)(N+1)^2} . \quad (4.28)$$

We have, using basic properties:

$$\sum_{p=1}^m \text{Var}[R_N^{<p>}(U)] = \frac{(N+1)m(m+1) - 2/3m(m+1)(2m+1)}{(N+2)(N+1)^2} , \quad (4.29)$$

Density	Analytical form	Support $\Omega(\cdot)$	Range
Uniform	$f_U(\zeta) = \frac{1}{2\sqrt{3}}$	$[-\sqrt{3}, \sqrt{3}]$	$2\sqrt{3}$
Linear	$f_L(\zeta) = \frac{\zeta + \sqrt{8}}{9}$	$[-\sqrt{8}, \sqrt{8}/2]$	$3/2\sqrt{8}$
Triangular	$f_T(\zeta) = \frac{\zeta + \sqrt{6}}{6}$ ($\zeta < 0$) $\frac{\sqrt{6} - \zeta}{6}$ ($\zeta > 0$)	$[-\sqrt{6}, \sqrt{6}]$	$2\sqrt{6}$
“V”-shape	$f_V(\zeta) = -\zeta/2$ ($\zeta < 0$) $f_V(\zeta) = +\zeta/2$ ($\zeta > 0$)	$[-\sqrt{2}, \sqrt{2}]$	$2\sqrt{2}$

Table 4.1. Four unit-variance random variables: their pdf, support and range (note: $f_X(\zeta) = 0$ if $\zeta \notin \Omega(X)$). The pdf and cdf are plotted in Fig.4.7. where the supports are scaled to be included in $(0, 1)$.

and

$$\sum_{1 \leq i < j \leq m} \text{Cov}[R_N^{(i)}(U), R_N^{(j)}(U)] = \frac{m(m-1)}{6(N+2)(N+1)^2} \times \{-3m^2 + m(2N-3) + 2N\} .$$

Some algebraic manipulations lead to:

$$\text{Var}[\langle R_N^{(m)}(U) \rangle] = \frac{-3m^3 + 2m^2(N-2) + 3mN + (N+1)}{3m(N+2)(N+1)^2} . \quad (4.30)$$

The theoretical curves showing $E[\langle R_N^{(m)}(U) \rangle]$ and $\text{Var}[\langle R_N^{(m)}(U) \rangle]$ (respectively given by Eq. (4.25) and Eq. (4.30)) are plotted in Fig. 4.9. They are compared to their empirical counterparts $E_t[\langle R_N^{(m)}(U) \rangle]$, $\text{Var}_t[\langle R_N^{(m)}(U) \rangle]$ where the subscript t means that the quantities are estimated via empirical mean/variance with t trials. They perfectly match.

- Joint effect of $p_X(x)$ and m

Because in ICA contexts one often deals with white signals, we shall consider here the whitened versions of U, L, T, V (the notation are the same as for those with range in $[0, 1]$ in order to avoid defining yet new symbols; the context should avoid confusion). The ranges of these white signals are given in Table 4.1. The samples are built by using the “sampling via inversion of the cdf” method, based on the property that the cdf $P_X(x)$ is uniformly distributed on $(0, 1)$ [Feller, 1966].

Figure 4.9. shows the empirical mean E_t and variance Var_t over t trials of $\langle R_N^{(m)}(\cdot) \rangle$ as a function of m . The variance of the estimator decreases with m for linear and triangular random variable, while this behavior cannot be observed for the white random variable U or V ; the variance of the

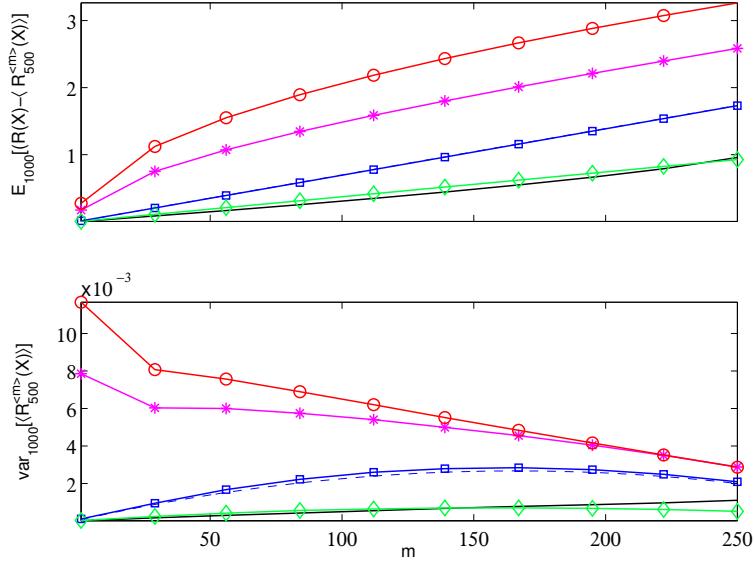


Figure 4.9. Empirical expectations $E_{1000}[\mathcal{E}_{500}^{<m>}(X)]$ (top) and variances $\text{Var}_{1000}[\langle R_{500}^{<m>}(X) \rangle]$ (bottom) as a function of m for various rv X . The theoretical curves for the uniform case are shown in dashed. Legend: uniform ('□'), linear('*'), triangular ('○') and "V"-shape ('◊') whitened variables.

estimators increases when unreliable points (i.e. corresponding to a low value of the pdf) are taken into account. The shape of $\text{Var}_t[\langle R_m(U) \rangle]$ is more surprising, but has been confirmed by the analytical equations given in Eq. (4.30).

- Performances comparison on Noisy Observations

The above results show that $m = 1$ leads to a good choice in the sense that obviously, the bias is minimized for this value of m . In addition, for a given N , $E[\mathcal{E}^{<m>}(U)]$ and $\text{Var}[\mathcal{E}_N^{<m>}(U)]$ both increase with m .

But this is in the noise-free case. In the case of unbounded additive Gaussian noise, the *polluted* random variable is noted $U^n \doteq U + G_n$ where G_n is Gaussian with variance σ_n^2 given by the SNR $10 \log \frac{\text{Var}[U]}{\sigma_n^2}$ (i.e. $-20 \log \sigma_n$ when white rv are considered). When G_n is fixed, $E_t[\langle R_N^{<m>}(U) \rangle]$ and $\text{Var}_t[\langle R_N^{<m>}(U) \rangle]$ decrease with m , and respectively tend to the theoretical values $E[\langle R_N^{<m>}(U) \rangle]$ and $\text{Var}[\langle R_N^{<m>}(U) \rangle]$ obtained without noise. This is shown on Figure 4.10. for a SNR equal to 35 dB.

A similar asymptotic behavior is expected for $\langle R_N^{<m>}(X) \rangle$ whatever the density p_X , even if we are not able to compute the exact expectations and variances. Therefore, the higher m , the lower the noise effect on the expectation and variance of $\mathcal{E}_N^{<m>}(X)$. As a conclusion, in noisy environment,

the value of m must be kept “quite small”; there is a tradeoff. The parameter m must be chosen neither too small (upper-estimation of the range and high variance due to noise), nor too high (the noise effect is cancelled but the theoretical noise-free bias increases with m).

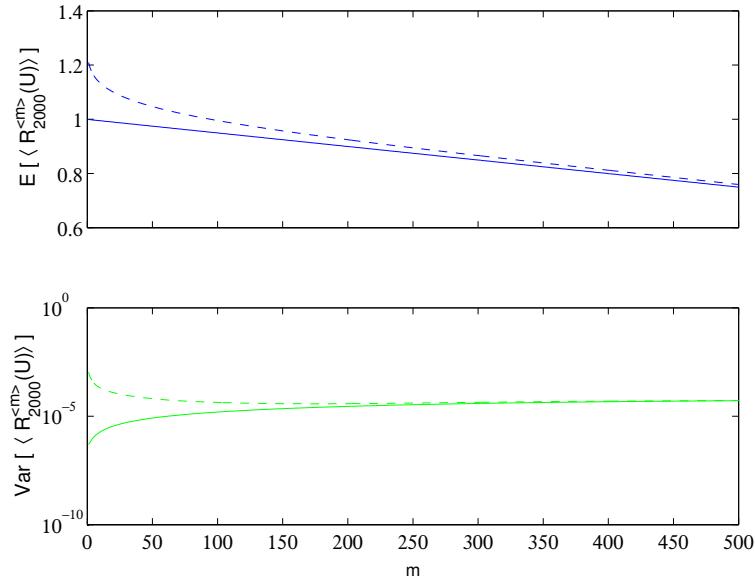


Figure 4.10. Top: theoretical expression of $E[\langle R_{2000}^{<m>}(U) \rangle]$ where U is a rv with support $\Omega(U) = [0, 1]$ (solid) and its empirical counterpart $E_{200}[\langle R_{2000}^{<m>}(U^n) \rangle]$ where U^n is the same as U but with additive Gaussian noise (dashed). Bottom: theoretical $\text{Var}[\langle R_{2000}^{<m>}(U) \rangle]$ (solid) and empirical $\text{Var}_{200}[\langle R_{2000}^{<m>}(U^n) \rangle]$ variances (dashed) (log scale). The signal-to-noise ratio involved in the noisy data set is 35 dB.

4.2.3.3 Impact on robustness of minimum-support ICA algorithms

We shall compare the effect of several parameters (namely: the sample size, the source density and the signal-to-noise ratio) on the robustness of minimum-support approaches to ICA when $\langle R_N^{<m>}(\cdot) \rangle$ is used as an estimator of the true range $R(\cdot)$. This is actually the most relevant viewpoint to test the quality of our estimator; for instance, even if the bias is high but constant with respect to θ , minimizing the estimated range would give the original sources. In our BSS framework, the robustness of a range estimator is viewed as the probability to induce a spurious solution by using the above range estimator in minimum range-based ICA methods; i.e. the lower the probability to face a spurious solution, the more robust the estimator.

For this study, we assume $K = 2$ and i.i.d. white and bounded sources, for simplicity purposes. The transfer vector is noted \mathbf{w}_θ leading to the output $\mathbf{Y}_\theta = \mathbf{w}_\theta \mathbf{S}$.

- Impact of $p_X(x)$

We start by considering the $m = 1$ case before discussing the joint impact of $p_X(x)$ and m on the quality of the range estimator for BSS.

The objective function $R(\mathbf{Y}_\theta)$ requires an estimation of the support width based on sample observations of \mathbf{Y}_θ only. One of the simplest estimations of $R(\mathbf{Y}_\theta)$ is the statistical range $R_N(\mathbf{Y}_\theta) = \langle R_N^{<1>}(\mathbf{Y}_\theta) \rangle$.

Figure 4.11. shows that the support estimation quality decreases when i) the source density is low near the bounds, and ii) the sources are more and more “mixed” (θ increases in from 0 to $\pi/4$).

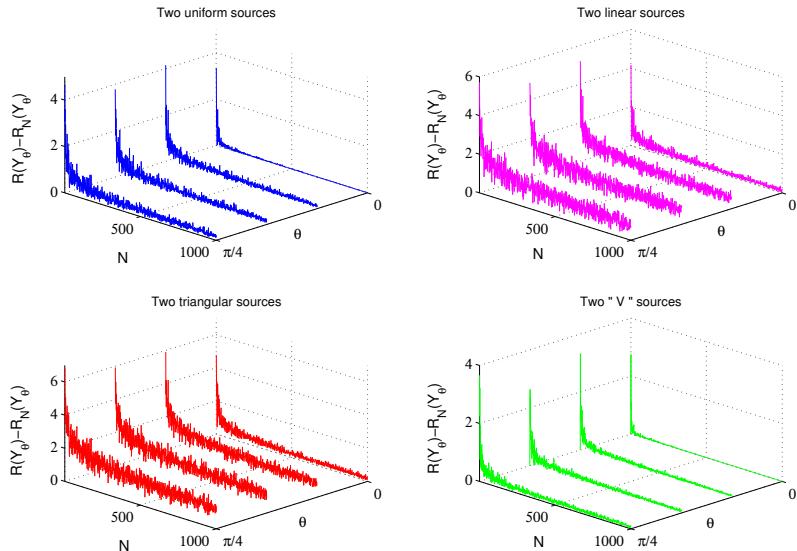


Figure 4.11. $\mathcal{E}_N(\mathbf{Y}_\theta)$ vs N and θ when \mathbf{S}_1 and \mathbf{S}_2 are two white uniform, linear, triangular and "V" white random variables.

Actually, both effects result from the same phenomenon. The density of the sum of two independent random variables is the convolution of their densities. Then, the density of \mathbf{Y}_θ varies with θ , and $p_{\mathbf{Y}_\theta}$ has less and less points in the neighborhood of its bounds when θ moves from 0 to $\pi/4$ (as an illustration, remind that the convolution of two rectangles is a trapeze or a triangle). This is illustrated by the histograms of Fig. 4.12.

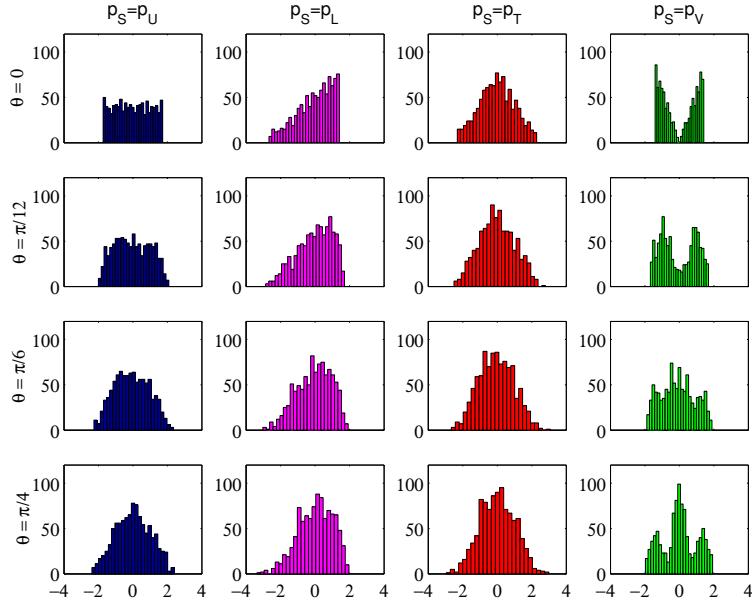


Figure 4.12. Samples histograms of Y_θ when both white sources follow a uniform, linear, triangular or a "V"-shape density.

The support estimation approach to ICA can give spurious solution, since if only few points are available, the estimated range of $Y_{(2k+1)\pi/4}$ can be lower than the one of $Y_{k\pi/2}$ (even if this is not possible when considering the true range itself). Then, the minimum support approach could suggest $Y_{(2k+1)\pi/4}$ as an estimate of one of the source, i.e. the algorithm is totally misled. If there are not enough sample points (i.e. N too small), and if the source pdf have a low density near its bounds w.r.t. the density near the center of the pdf (as is e.g. the case for a triangular pdf), then we could have $R_N(Y_{\pi/4}) - R_N(Y_{k\pi/2}) < 0$ with $k \in \mathbb{Z}$ (see Fig. 4.13.)

The following simulation results give a final illustration. Again, we can observe in Fig. 4.13. that $R_N^{<m>}(Y_\theta)$ is a bad estimate of the true support when i) θ is very different from the closest $k\pi/2$ angle, ii) when the number of sample points is small, and iii) when the source pdf is concentrated around its mean.

As a conclusion, Figure 4.14. shows that the minimum-support method can be applied to noiseless BSS with the observed range estimator, but the required number of sample points to have satisfactory performances and to avoid spurious minima depends on the last pdf.

What about another choice of m ? Figures Fig. 4.15. and Fig. 4.16. show $E_t[\langle R_N^{<m>}(Y_\theta) \rangle]$ and $\text{Var}_t[\langle R_N^{<m>}(Y_\theta) \rangle]$ for various values of m ; the usual

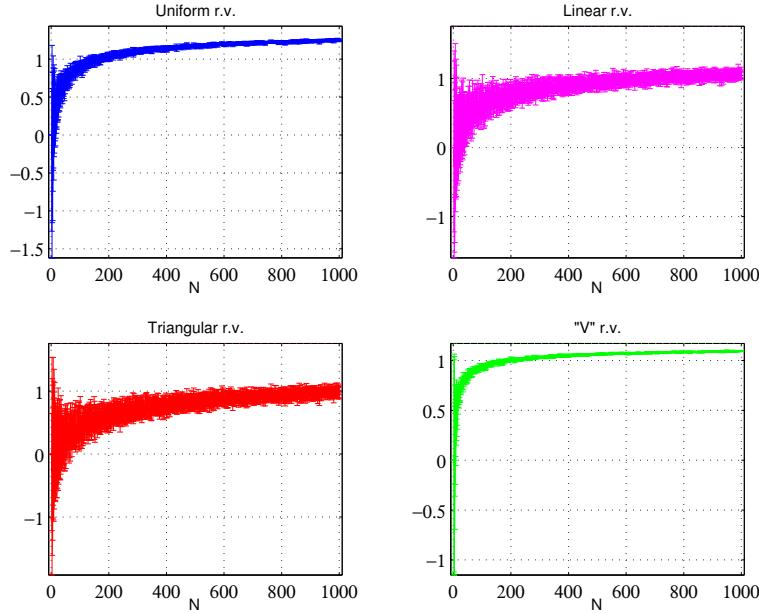


Figure 4.13. $E_{50}[R_N(Y_{\pi/4}) - R_N(Y_{k\pi/2})]$. The width of the curve is $2\text{Var}_{50}[R_N(Y_{\pi/4}) - R_N(Y_{k\pi/2})]$, and the curve tend to the upper side of the bounding box as $N \rightarrow \infty$. We observe that one can face $R_N(Y_{\pi/4}) - R_N(Y_{k\pi/2}) < 0$ when simultaneously N is too small and the common source pdf has a low density near the bounds.

concave shape is lost when m is too large compared to N . Even the worst case can occur. Assume that we have two white uniform sources. For $\theta = \pi/4$, the output Y_θ has a triangular density, and this output still has a unit-variance. The hot point is that it is feared from Figure 4.9. that for sufficiently large m compared to N , the range of the unit-variance triangular r.v. can be underestimated in a so strong way (compared to the error made on the range estimation of a white uniform r.v.) that the estimated range $\langle R_N^{<m>} \rangle$ of a white triangular r.v. can be lower than the estimated range of a white uniform source; clearly this is an aberration as the true range of a white triangular density (which equals $2\sqrt{6}$) is always larger than the true range of a white uniform rv (which equals $2\sqrt{2}$). The estimator however, could yield such a paradoxical value. Indeed this phenomenon can be observed for $m \approx 100$ if $N = 500$. The best choice of m is thus very difficult: neither too small (large variance), neither to high (spurious minima at $\theta = (2k + 1)\pi/4$).

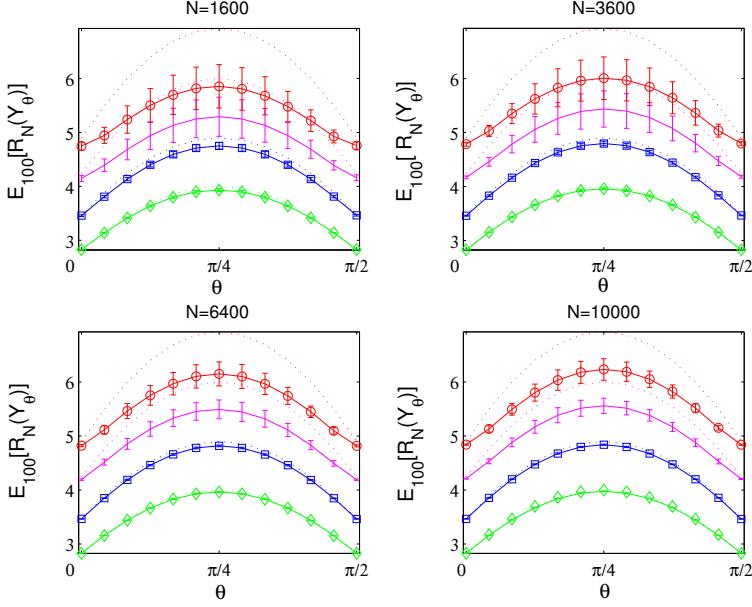


Figure 4.14. $E_{100}[\langle R_N^{<m>}(Y_\theta) \rangle]$ vs θ for $m = 1$. We observe a rather low bias for a given N , rounded shape as desired, but high variance between trials (low confidence). Legend: pair of uniform (' \square '), triangular (' \circ '), "V"-shape (' \diamond ') and linear (no marker) sources.

- On the choice of m

Is the choice of m critical? In any case, it does matter, so how can we choose the value of m ? The quantity $\mathcal{E}_N^{<m>}$ increases with m , at a rate depending on the density p_X . In order to have m so that $\langle R_N^{<m>}(X) \rangle$ is still a “relevant image” of the support (from the BSS point of view), m must be small enough, but not equal to one, in order to save robustness. In this section, we propose a method to choose a value for m . The main idea is to fix m such that the error $\mathcal{E}_N^{<m>}(X)$ is lower than an error threshold ϵ_τ with a high probability, whatever the density of X . In other words, we try to find m_0 such that for all $m \leq m_0$:

$$\Pr [\mathcal{E}_N^{<m>}(X) \leq \epsilon_\tau] \geq L(m_0) , \quad (4.31)$$

where $L(m_0)$ is a probability threshold ideally close to, but lower than one. The problem is that if ϵ_τ is constrained to be a constant, the latter probability depends on $P_{R_N^{<m>}}$ (that is on p_X), which is supposed to be unknown here. The trick consists in choosing ϵ_τ of the form $R(X) - (\xi_\alpha - \xi_\beta)$, where ξ_α is the α -th quantile of P_X , i.e. $P_X(\xi_\alpha) = \alpha$, with $0 \leq \alpha \leq 1$.

By doing so, the value of ϵ_τ cannot be found explicitly without knowing the density of P_X , but it can be made small by taking α close to one and

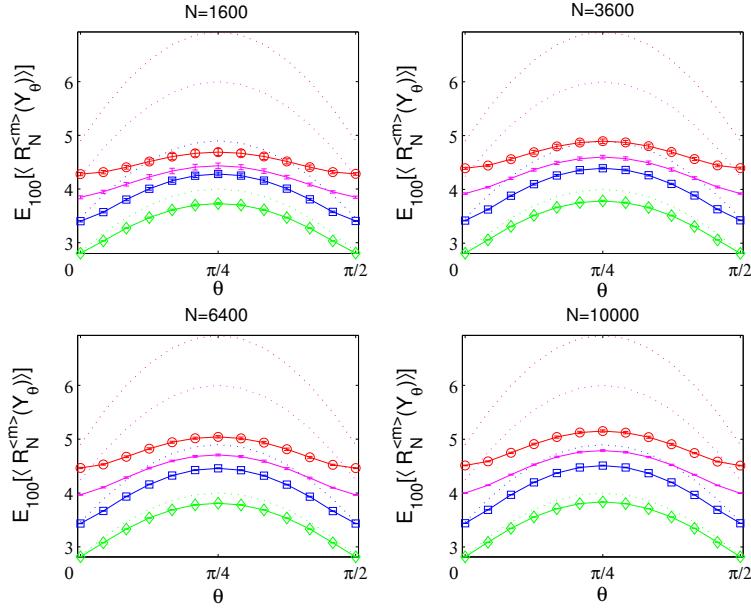


Figure 4.15. $E_{100}[\langle R_N^{<m>}(Y_\theta) \rangle]$ vs θ for $m = \sqrt{N/2}$. We observe a higher bias for a given N , acceptable but flatter shape, low variance between trials (better confidence). Legend: pair of uniform (' \square '), triangular (' \circ '), "V"-shape (' \diamond ') and linear (no marker) sources.

β close to zero. In order to find a suitable bound L , we start from an inequality due to Chu [Chu, 1957]:

$$\begin{aligned} \Pr[X_{(s:N)} - X_{(p:N)} &\geq \xi_\alpha - \xi_\beta] \geq \sum_{i=p}^N \binom{N}{i} \beta^i (1-\beta)^{N-i} \\ &\quad - \sum_{i=s}^N \binom{N}{i} \alpha^i (1-\alpha)^{N-i} \end{aligned} \quad (4.32)$$

Let us choose $\beta = 1 - \alpha$, $\alpha > 1/2$ and set $s = N - m_0 + 1$, $p = m_0$; this implies that

$$\Pr[R(X) - R_N^{<m_0>}(X) \leq \underbrace{R(X) - (\xi_\alpha - \xi_{1-\alpha})}_{\epsilon_\tau}] \quad (4.33)$$

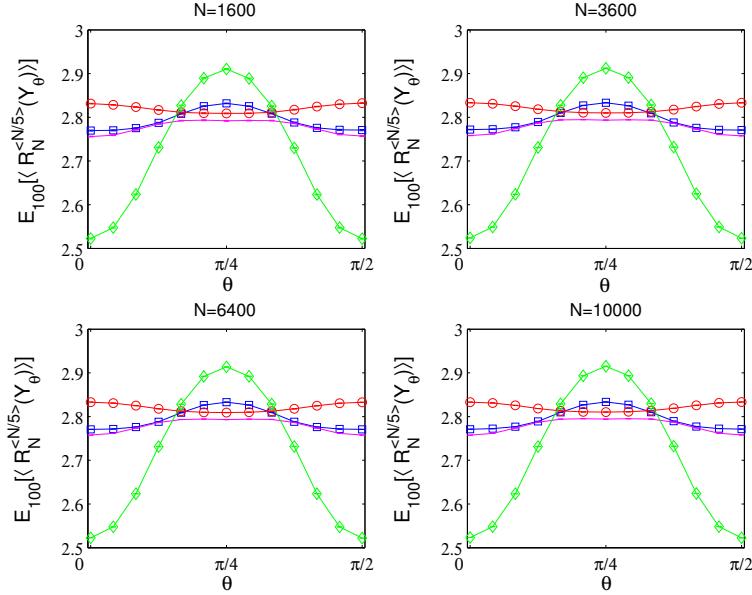


Figure 4.16. $E_{100}[\langle R_N^{<m>}(Y_\theta) \rangle]$ vs θ for $m = N/5$. High bias for a given N , dangerous shape, but extremely low variance between trials (higher confidence). Legend: pair of uniform (' \square '), triangular (' \circ '), "V"-shape (' \diamond ') and linear (no marker) sources.

is lower-bounded by

$$L(\alpha, m_0, N) \doteq \sum_{i=m_0}^N \binom{N}{i} (1-\alpha)^i \alpha^{N-i} - \sum_{i=N-m_0+1}^N \binom{N}{i} \alpha^i (1-\alpha)^{N-i}. \quad (4.34)$$

Note that for large numbers, numerical problems may occur when computing $\binom{N}{i}$. Therefore, it is recommended to use the following logarithmic trick:

$$\binom{N}{p} = e \left[\sum_{i=1}^N \log i - \sum_{i=1}^{N-p} \log i - \sum_{i=1}^p \log i \right]. \quad (4.35)$$

It is obvious that $L(\alpha, m_0, N)$ is also a lower bound on $\Pr[\mathcal{E}_N^{<m_0>}(X) \leq \epsilon_\tau]$ whatever the density of the random variable X . Indeed, any lower bound on $\Pr[R_N^{<m_0>}(X) \geq R(X) - \epsilon_\tau] = \Pr[R(X) - R_N^{<m_0>}(X) \leq \epsilon_\tau]$ can be

used as a lower bound on: $\Pr[R(X) - \langle R_N^{<m_0>}(X) \rangle \leq \epsilon_\tau]$:

$$\begin{aligned}
\Pr[\mathcal{E}_N^{<m_0>} \leq \epsilon_\tau] &= \Pr[\langle R_N^{<m_0>}(X) \rangle \geq R(X) - \epsilon_\tau] \\
&= \Pr[\langle R_N^{<m_0>}(X) \rangle \geq R(X) - \epsilon_\tau | R_N^{<m_0>}(X) \geq R(X) - \epsilon_\tau] \\
&\quad \times \Pr[R_N^{<m_0>}(X) \geq R(X) - \epsilon_\tau] \\
&+ \Pr[\langle R_N^{<m_0>}(X) \rangle \geq R(X) - \epsilon_\tau | R_N^{<m_0>}(X) < R(X) - \epsilon_\tau] \\
&\quad \times \Pr[R_N^{<m_0>}(X) < R(X) - \epsilon_\tau] \\
&\geq \Pr[R_N^{<m_0>}(X) \geq R(X) - \epsilon_\tau], \tag{4.36}
\end{aligned}$$

where the inequality results from the fact that $\langle R_N^{<m_0>}(X) \rangle \geq R_N^{<m_0>}(X)$ with probability one.

The parameter m is thus chosen once α (controlling the lower bound on $\langle R_N^{<m>}(X) \rangle$) and $L(\alpha, m_0, N)$ (controlling the probability that $\langle R_N^{<m>}(X) \rangle$ is greater than $\xi_\alpha - \xi_{1-\alpha}$) are fixed, ideally close to one. We choose $m = m_0$, i.e. the largest value of m such that $\langle R_N^{<m_0>}(X) \rangle \geq (\xi_\alpha - \xi_{1-\alpha})$ with a probability higher than a threshold P_τ . Both P_τ and α have to be fixed, and m comes by finding the largest value m_0 such that $L(\alpha, m_0, N) \geq P_\tau$ holds. Figure 4.17.(a) shows the area in the (m, N) space where the bound L is useful (i.e. strictly positive).

Figure 4.17.(b) shows $L_+(0.95, m, N)$ where

$$L_+(\alpha, m, N) = \max(0, L(\alpha, m, N)) \tag{4.37}$$

(negative probability bounds are useless) as a function of (m, N) . We see that m_0 is given by finding the maximum value of m such that the inequality $L(0.95, m, N) \geq P_\tau$ holds, with fixed P_τ , e.g. $P_\tau = 90\%$. The value m^\sharp is a value given by the rule of the thumb

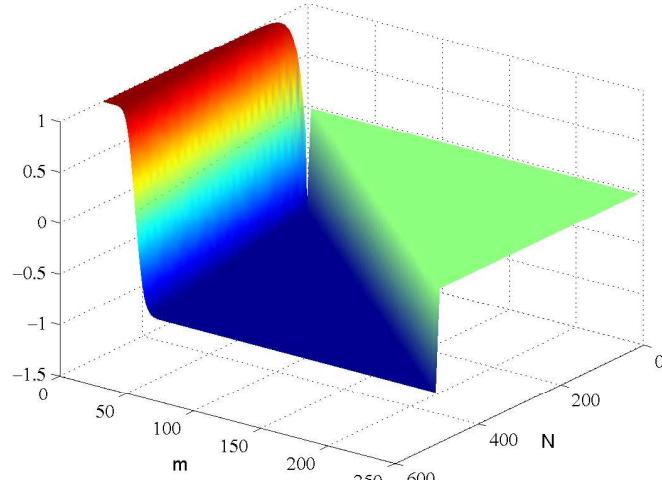
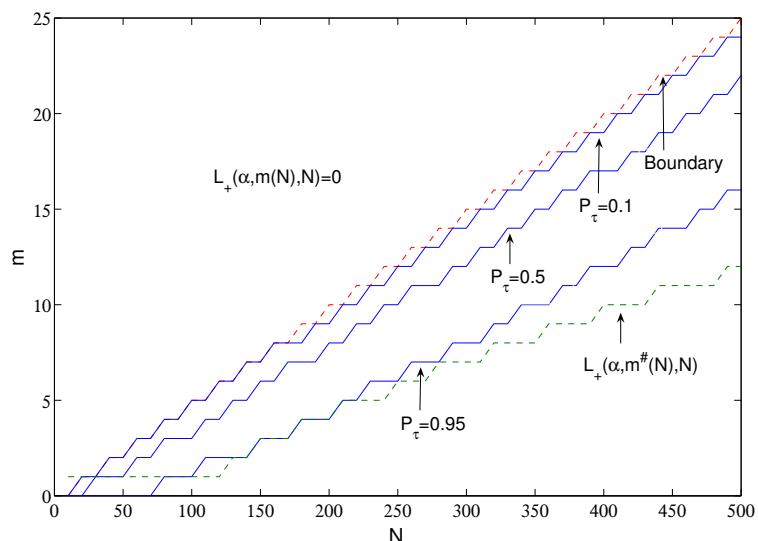
$$m^\sharp(N) \doteq \max \left(1, \Re \left(\left[\overline{\left(\left(\frac{N-18}{6.5} \right)^{0.65} - 4.5 \right)} \right] \right) \right), \tag{4.38}$$

where $\overline{\alpha}$ denotes the nearest integer to α , which aims at approximating the largest value m_0 of m such that $L(0.95, m, N) \geq .95$. This highly improves the computational requirement for finding a suitable value of m .

Part of the above results were published [Vrins and Verleysen, 2006a,b].

4.3 RANGE MINIMIZATION ALGORITHM: SWICA

Usually, when one desires to optimize a criterion for which an algebraic solution cannot be found, iterative methods are used, as explained in Section 1.6.

(a) $L(\alpha, m, N)$ vs (m, N) .(b) Selected iso- $L_+(\alpha, m, N) = P_\tau$ curves for various probability thresholds P_τ . The curve $L_+(\alpha, m^*(N), N)$ given by Eq. (4.38) vs N has also been plotted. This bound is useless ($L(\alpha, m, N) \leq 0$) above the "Boundary" curve.**Figure 4.17.** Dependency of the probabilistic error bound on the range estimation vs (m, N) , $\alpha = 0.95$.

When gradient-ascent methods are plugged in BSS methods, the obtained solution is given by the stationary point corresponding to the local maximum of the criterion. The problem is that when considering the range-based contrast function, the local maximum points are not stationary points of the gradient-ascent update rule as the gradient does not exist at these points. Therefore, iterative maximization schemes have to be developed for such kind of criteria. This is precisely the aim of this section: a specific algorithm for the optimization of non-differentiable contrasts is proposed. The algorithm actually performs a geodesic optimization of the BSS criterion over the manifold of the orthogonal matrices. It is proved to succeed in maximizing the range-based D-BSS contrast $\mathcal{C}_R(\mathbf{b})$. This algorithm was first proposed in [Lee et al., 2005]. Note that even simultaneous and partial counterparts may be easily found based on the theoretical criteria, we focus our experiments on deflation extraction schemes as it seems that they yield more interesting results even in high-dimensional spaces.

4.3.1 Algorithm

The simple algorithm shown in Table 4.2. may be used to compute each row of \mathbf{B} . Briefly, the so-called “ICAforNDC” algorithm looks at a given contrast $\mathcal{C}(.)$ value in some (mutually perpendicular) directions.

It could be interesting to use prewhitening in order to reduce the dimension of the search space, as previously explained. In other words, we would like to perform a geodesic optimization over the group of orthogonal matrices.

How to modify the \mathbf{b}_i 's while not affecting the orthogonality of \mathbf{B} ? As each \mathbf{b}_i is orthogonal to any other row \mathbf{b}_j , any linear combination of the form $\cos(\alpha) \mathbf{b}_i + \sin(\alpha) \mathbf{b}_j$ will be orthogonal to any row \mathbf{b}_k of \mathbf{B} where $k \notin \{i, j\}$. Obviously, this new value of \mathbf{b}_i is no longer orthogonal to \mathbf{b}_j . A new value for \mathbf{b}_j which is orthogonal to any other row of \mathbf{B} is given by $\cos(\alpha) \mathbf{b}_j - \sin(\alpha) \mathbf{b}_i$ (it can be checked that the dot product between the new values of $\mathbf{b}_i, \mathbf{b}_j$ is zero).

For short, we can note the above positive and negative angular variations of \mathbf{b}_i as

$$\mathbf{b}_{i \uparrow j} = \cos(\alpha) \mathbf{b}_i + \sin(\alpha) \mathbf{b}_j , \quad (4.39)$$

$$\mathbf{b}_{i \downarrow j} = \cos(\alpha) \mathbf{b}_i - \sin(\alpha) \mathbf{b}_j , \quad (4.40)$$

where α is a kind of “learning parameter”, which value (decreasing between two iterations) controls the “amount of the variation”. Then, the orthogonality of \mathbf{B} is preserved if \mathbf{b}_i is replaced by $\mathbf{b}_{i \uparrow j}$ provided that \mathbf{b}_j is replaced by $\mathbf{b}_{j \downarrow i}$.

Observe that these update rules can be jointly generated by left-multiplying the current demixing matrix by the Givens matrix: $\mathbf{B} \leftarrow \mathbf{G}_{ij}^\alpha \mathbf{B}$ (see Section 1.6). In this case, because of the group structure of $\mathcal{SO}(K)$ and $\mathbf{G}_{ij}^\alpha \in \mathcal{SO}(K)$, the algorithm actually performs a geodesic optimization on $\mathcal{SO}(K)$ if the initial demixing matrix is in this subset. More precisely, the algorithm allows us, at each step, to explore the contrast in $\mathcal{SO}(K)$, but only in specific (pairwise orthogonal) planes spanned by the pair of row-vectors $(\mathbf{b}_i, \mathbf{b}_j)$ (i.e. only along Jacobi trajectories).

$[\mathbf{BV}] = \text{ICAforNDC}(\mathcal{C}, \mathbf{X}(t))$

1. Whiten the mixtures using an eigenvalue value decomposition:
 - (a) remove the mean: $\mathbf{X}(t) \leftarrow \frac{1}{N} \sum_{t=1}^N \mathbf{X}(t)$
 - (b) compute the whitening \mathbf{V} matrix of $\mathbf{X}(t)$ by using Eq. (1.47)
 - (c) compute the projected whitened mixtures: $\mathbf{X}(t) \leftarrow \mathbf{V}\mathbf{X}(t)$
 2. Initialize $\mathbf{B} \leftarrow \mathbf{I}_K$.
 3. To extract the i -th source, with $1 \leq i \leq K$, do, for k ranging from 1 to 50:
 - (a) $\alpha \leftarrow \pi 0.75^k$.
 - (b) For j ranging from $i + 1$ to K , determine the best contrast value:
 - if $\mathcal{C}(\mathbf{b}_{i\downarrow j}) > \mathcal{C}(\mathbf{b}_i)$ and $\mathcal{C}(\mathbf{b}_{i\downarrow j}) > \mathcal{C}(\mathbf{b}_{i\downarrow j})$ then $\mathbf{b}_i \leftarrow \mathbf{b}_{i\downarrow j}, \mathbf{b}_j \leftarrow \mathbf{b}_{j\downarrow i}$
 - else if $\mathcal{C}(\mathbf{b}_{i\downarrow j}) > \mathcal{C}(\mathbf{b}_i)$ and $\mathcal{C}(\mathbf{b}_{i\downarrow j}) > \mathcal{C}(\mathbf{b}_{i\downarrow j})$ then $\mathbf{b}_i \leftarrow \mathbf{b}_{i\downarrow j}, \mathbf{b}_j \leftarrow \mathbf{b}_{j\uparrow i}$
 - end.
 4. return \mathbf{BV}
-

Table 4.2. Pseudo-code for the deflation ICA algorithm for non-differentiable contrast functions $\mathcal{C}(\cdot)$.

The corresponding contrast values can be written as $\mathcal{C}(\mathbf{b}_{i\downarrow j})$ and $\mathcal{C}(\mathbf{b}_{i\uparrow j})$ (remember that the mixtures are supposed to be whitened).

When the learning parameter is set to the default value of 0.75, the algorithm usually converges after ten or twenty iterations; default is 50 for security.

By construction, the algorithm is monotonic: the contrast is either increased or kept constant. For this reason, if spurious maxima of the contrast exist, then the algorithm could fall in one of them, especially if the initial values of α are too small or if α decreases too fast during the first iterations.

When the range-based D-BSS contrast $\mathcal{C}_R(\mathbf{b})$ is used in the algorithm, it is referred to as SWICA (the acronym stands for “Support Width ICA”).

Remark 22 *The first goal of this thesis was not to develop optimization schemes but rather to analyze the theoretical behavior of entropic contrast. Therefore, a wide comparison of optimization techniques of non-differentiable functions has not been performed. However, other algorithms using more sophisticated techniques have been tried too. Unfortunately, they lead to worse results than the proposed methods both in terms of computational complexity and separation performances. In addition, they involve a larger number of parameters, which are tedious for adjusting. We are convinced that the speed of the contrast maximization schemes presented below could be largely improved by investigating other optimization tools, such as sub-differential techniques [Erdogan, 2006]. Nevertheless, in spite of their (probably) sub-optimal convergence rates, the simple proposed algorithms would deserve attention as they yield very promising results*

in terms of separation performance indexes compared to other general-purpose separation algorithms in some situations.

Another fact that deserves to be emphasized is that Jacobi-like algorithms (like ICAforNDC) may be stuck in spurious optima, even when the (possibly non-differentiable) contrast function is discriminant. This is e.g. the case of the range criterion as explained in Section 3.4.5 (probably due to the fact that the function is not differentiable everywhere). In particular, Jacobi-like algorithms can yield spurious solutions even for piecewise g-convex contrast function, as defined in Section 3.4.2.2.

4.3.2 Performance analysis of SWICA for OS-based range estimators

In this section, we compare the extraction performances of 3 ICA algorithms: FastICA (developed in [Hyvärinen and Oja, 1997]), JADE (introduced by [Cardoso, 1989, Cardoso and Souloumiac, 1993]), and SWICA with the following range estimator $R(\mathbf{X}) \approx \langle R_N^{<m>}(\mathbf{X}) \rangle$ (when this specific range estimator is plugged in the SWICA approach, the method is referred to as “AVOSICA”, for “sAveraged Order Statistics ICA”).

Note that other “SWICA” approaches involving different range-based estimators ($R(\mathbf{X}) \approx R_N^{<m>}(\mathbf{X})$ or $R(\mathbf{X}) \approx R_N^{<1>}(\mathbf{X}) = \langle R_N^{<1>}(\mathbf{X}) \rangle$) have been tested too, but they are not shown here as they lead to worse results.

The default value for the parameter m was chosen equal to m^\sharp , given by Eq. (4.38). The algorithms have been tested on the extraction of 5 bounded and white sources from 5 mixtures (the source densities are given in Table 4.1.). The mixing matrix is built from 25 random coefficients uniformly distributed in $(0, 1)$.

Figure 4.18. compares the histograms of the SPI (Square Performance Index) for each extracted source in the noise-free case for $N = 2000$ and $m = m^\sharp(N)$:

$$\text{SPI}(\mathbf{S}_i) \doteq \frac{\sum_{j=1}^n W_{ij}^2}{\max_j W_{ij}^2} - 1 . \quad (4.41)$$

The global SPI is defined as the average of the $\text{SPI}(\mathbf{S}_i)$ over i (i.e. over the sources). A zero $\text{SPI}(\mathbf{S}_i)$ indicates that \mathbf{Y}_i is proportional to a source, while a high $\text{SPI}(i)$ means that \mathbf{Y}_i results from the superimposition of several sources.

We can observe in Figure 4.18. that AVOSICA gives the most interesting results, in comparison to JADE and FastICA (pow3), especially for the separation of sources with linear and triangular pdf. It must be stressed that even if AVOSICA performs quite satisfactorily for small values of N , the performances are improved for large N .

Figure 4.19. summarizes the global SPI performances of ICA algorithms for various noise levels. Note that the performance results are analyzed from the mixing matrix recovery point of view; the source denoising task is not considered here. The good results of AVOSICA can be observed, despite the fact that the value of m has not been chosen to optimize the results, i.e. we always have taken $m = m^\sharp(N)$. It must be stressed that the value of the parameter m is not critical when chosen around $m^\sharp(N)$.

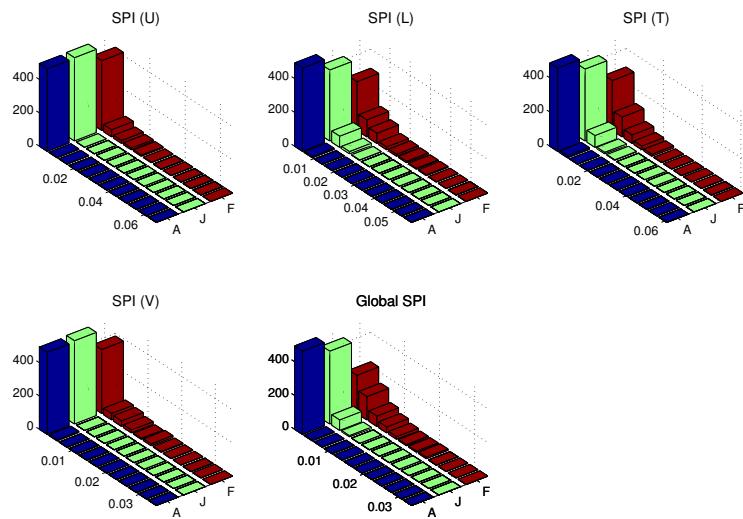


Figure 4.18. 12-bins histograms of SPI for each extracted source, for 100 trials, $N = 2000$, and $m = m^\sharp(N) = 37$. The analyzed algorithms are AVOSICA ('A'), JADE ('J') and FastICA ('F'). The *global SPI* is the averaged SPI computed from the individual source SPIs for a given trial.

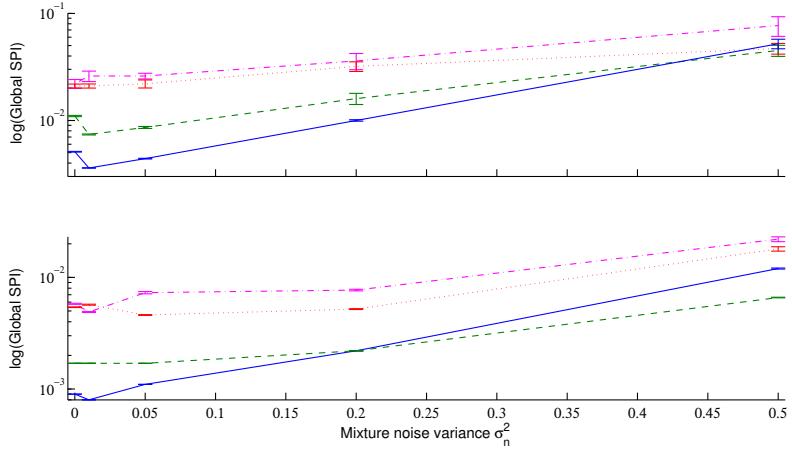


Figure 4.19. 100-trials-averaged empirical means and variances of global SPI performances of several ICA algorithms as a function of additive Gaussian noise with standard deviation σ_n : $N = 500$ (top panel) and $N = 2000$ (bottom panel). Legend: AVOSICA (solid), JADE (dashed), FastICA-pow3 (dotted), FastICA-gauss (dash-dotted). The noise has been added to the whitened mixtures, so that for a given σ_n , the “mixture SNRs” equal $-10 \log \sigma_n^2$; the noise variance does not vary between trials, and does not depend of the mixing weights). For AVOSICA, the parameter is $m = m^\sharp(N)$.

JADE is a very good alternative when the dimensionality of the source space is low. The computational time of FastICA is its main advantage.

Part of these results appeared [Vrins and Verleysen, 2006b].

4.4 EXTENSIONS AND APPLICATIONS OF THE MINIMUM RANGE APPROACH TO ICA

It has been already mentioned that when the sources are truly independent, one can first apply a decorrelation transform to the mixtures, and then look for a linear transformation that preserves the covariance matrix yielding independent outputs. This makes sense because correlation is a linear dependency. But what happens if the sources are not perfectly independent (their mutual information is non-zero)? In this section, we present an application involving correlated sources.

The ICAforNDC algorithm (and in particular, SWICA) performs a geodesic optimization over the group of orthogonal matrices. In other words, SWICA performs rotations of the mixing-whitening matrix \mathbf{VA} . In some cases however, as explained below, one desires to find a demixing matrix which is not orthogonal,

even after a prewhitening step; this is e.g. the case when the source signals are correlated: \mathbf{VA} is no longer orthogonal. Even if the convex hull of the source scatter plot forms a rectangle (in spite of the correlation), the whitened mixture scatter plot convex hull forms a parallelepiped. Therefore, we are looking for a demixing matrix that maps a parallelepiped to a rectangle. To that end, we consider an extension of the SWICA algorithm, namely the NOSWICA method (where the acronym stands for Non-Orthogonal SWICA). The motivation and the method are described below. Under the source independence assumption, both yield the same solution, but SWICA performs the optimization in $\mathcal{SO}(K)$ while NOSWICA performs an optimization in $\mathbb{R}^{K \times K}$. Therefore, NOSWICA can be seen as a *least dependent component analysis* in which decorrelation (linear dependency) is not enforced; the “least dependent” point of view is obviously related to the range. This is explained below.

4.4.1 The problem of blind images separation : NOSWICA

4.4.1.1 Application of SWICA on correlated images separation

Images separation has proved to be a successful and realistic application of BSS [Almeida and Faria, 2004]. The most common example is the case where each face of a tiny sheet of paper is scanned. Both scans reflect the information of the corresponding face that has been scanned, but the information located on the other face also appears in the scan. Then, each of the scans is a mixture of two (assumed to be independent) images. In this work, we shall consider a toy example involving a linear mixture of images. It seems that non-linear mixing schemes are more realistic, but linear mixing schemes can be seen as first approximations of more complicated models. Furthermore, even if the linear mixture does not really correspond to reality, it emphasizes an interesting problem that is difficult to address with standard BSS algorithms, but that can be dealt with by using geometric-based method, such as SWICA.

When considering similar images such as two landscapes, two human face pictures (etc), they may seem to be independent (the pictures represent different landscapes, or different persons), but actually, the images are correlated. This is because the image can be divided in two parts: a “background part” and a “detail part”. For instance, two landscapes have a common shape: globally, they are built from horizontal shapes; two human face pictures shows a background and a “disk” representing the face. Of course, the precise landscapes and faces make that the pictures look different but globally, they share a same “template”. Hence, even if exceptions seem to exist [Yang and Amari, 1997], two mixed images can be more independent than the dependent sources; in particular, they can be decorrelated (linearly independent). Several tools have been designed to address this issue. The most efficient one seems to be filtering (frequency masking) [Cichocki and Georgiev, 2003]. In that case, it is assumed that a frequency band exists (corresponding to the “detail” part of the figure, as described above), in which the source images are statistically independent. By filtering the image

mixtures outside this frequency band, new mixtures are obtained and processed by a usual ICA algorithm. Once the latter has converged, the computed unmixing matrix is used for separating the initial (unfiltered) mixed images [Cichocki and Georgiev, 2003].

Even if the previous method looks very efficient, additional parameters appear, like the cutoff frequencies and the order of the filter, which may be difficult to adjust. For instance, finding the frequency band that makes unknown images fully independent, starting from mixtures of them may be a tedious task. To our knowledge, no simple or automatic method exists to solve this problem. Therefore, we suggest another way of solving this problem, based on the geometric nature of the images scatter plot that is precisely exploited by SWICA. This approach was presented in 2005 (see [Vrins et al., 2005c]).

The major limitation of range-based methods is that sources having pdf with flat tails are extremely difficult to extract, because the true range of the mixture is badly estimated, as explained in Section 4.2.3.3. This estimator works best for abruptly bounded variables. When the tails of the density are longer and less dense, as for platykurtic variables, the estimator may fail to give a good approximation of the support (see Fig. 4.20.(a)). The reason is that there are not enough observations in ‘critical areas’ (the ‘corners of the square’ in the figure). In the following of the section, SWICA refers to AVOSICA with $m = 1$, for short. SWICA may be completely misled: it has minimized the support of \mathbf{Y}_1 , but this output does not correspond to a source (we have set $m = 1$ to better point out the problem). Of course, if four well-chosen observations are available (located by the arrows in Fig. 4.20.(b)), the problem disappears. This shows

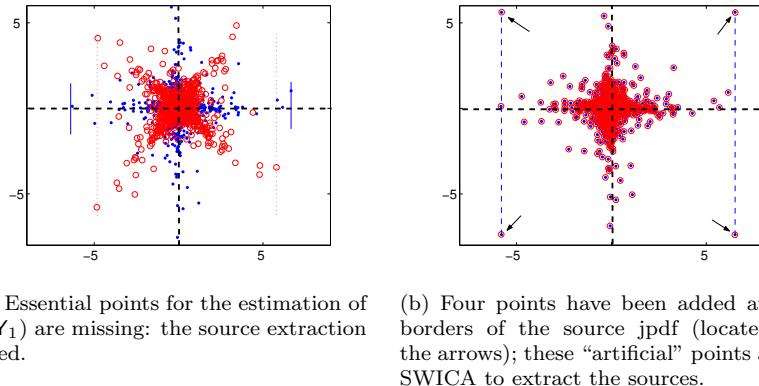


Figure 4.20. AVOSICA ($m = 1$) applied on super-Gaussian signals with $m = 1$: scatter plots of the source signals (dots) and of the outputs (circle).

once again that the AVOSICA approach is very sensitive to a small number of observations when m is small. But we remind it as we shall exploit this weakness later on. Fortunately, the densities of the pixel intensities in an image

are usually abruptly bounded, due to particular implementation choices (small encoding range) and image properties (because neighboring pixels often have similar values, there are usually few outliers). Observe, in addition, that only these four points are considering in the range estimation, whatever the mixing matrix (iff $m = 1$); the inner structure of the scatter plots does not play any role in the contrast evaluation. This will also be exploited in image separation, as shown in the next example.

Example 28 Consider the source images in Fig. 4.21.(a) and 4.21.(d); they are two gray-level landscapes whose 8-bits coded pixel intensities range from 0 to 255. A random linear mixing matrix and a decorrelation transformation are then applied; mixed images are shown in Fig. 4.21.(b) and 4.21.(e) (after translation and scaling to map the mixtures in the full range [0, 255], for readability purposes). The correlation between the images rises to 26%. This correlation is visible when looking at the scatter plot of the source images (Fig. 4.22.(a)); they are not independent because the joint probability density function cannot be factorized (for instance, look at several horizontal – or vertical – slices in the scatter plot (reflecting conditional densities): they are not equal to each other). Let us compare these results with FastICA and JADE. As could be feared, both fail to recover the source images (see Fig. 4.22.(b) and 4.22.(c)).

We shall use the SWICA deflation algorithm (remind that in this section, it is assumed that the range estimator is $R(\mathbf{X}) \approx R_N(\mathbf{X}) = \langle R_N^{<1>}(\mathbf{X}) \rangle$). The extraction of the first and second sources are considered separately.

- Extraction of the first source

SWICA behaves rather differently from JADE and FastICA and recovers one of the sources (the second one, see the corresponding estimated image in Fig. 4.21.(c) and the output scatter plot in Fig. 4.22.(d)). The output scatter plot is a parallelogram and two of its edges are parallel to the vertical axis (this is more clear for the well-drawn leftmost edge): values of \mathbf{Y}_1 computed by SWICA nearly equal those of \mathbf{S}_1 . Unfortunately, it is seen from Fig. 4.22.(d) that the two other edges of the parallelogram are not parallel to the horizontal axis, meaning that \mathbf{Y}_2 does not correspond to the yet un-extracted source (i.e. to the first one). Looking at the corresponding output image in Fig. 4.21.(f), one can see that the mountain of the second source image is not perfectly removed.

Understanding why both JADE and FastICA fail in this example is straightforward: because the source images are correlated, looking for (as independent as possible) outputs constrained to be uncorrelated is inadequate. In the case of two landscapes, these components are not the source images but new images (e.g. component 1 could account for the shared soil/sky contrast whereas component 2 could account for varying trees, mountain and clouds).

Contrarily to other ICA contrasts, the range criterion extracts a very limited piece of information out of the marginal pdf of the currently estimated

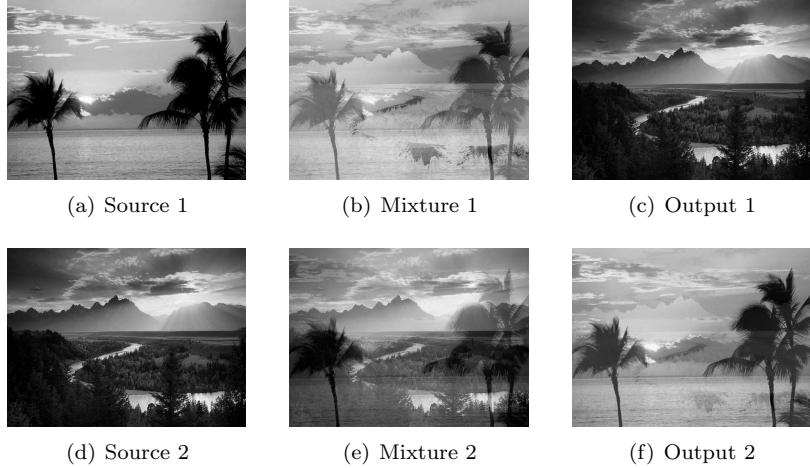


Figure 4.21. Example of images separation using SWICA ($\text{Cov}(\mathbf{S}_1, \mathbf{S}_2) = 0.26$); source images (a,d), rescaled mixed images (b,e), rescaled extracted images (c,f).

source \mathbf{Y}_1 : the bounds. In our image application, bounds are particularly interesting parts of the image density. Indeed, images may be assumed to involve three more or less important parts: (i) a global shared shape, (ii) local independent details and (iii) encoding techniques. The global shared shape leads to highly correlated and dense spots in the scatter plot. On the other hand, local independent details contribute to fill the scatter plot in a spread (“uniform”) but very sparse way. Finally, encoding techniques generally produce saturation effects (towards full white and/or black), which are independent (source images are independently encoded). Usual ICA contrasts are especially sensitive to the global shape, which is dominating in correlated images, and thus try to make the images independent. On the other hand, the range focuses on the bounds of the scatter plot: these bounds are generally well drawn due to parts (ii) and/or (iii) of the images and contain most of the independent features of the images. As an example, if there exist pixel indexes (that can be assumed to be uni-dimensional after a concatenation of the rows of the images) t_1, t_2, t_3, t_4 such that $\mathbf{S}(t_1) = [0, 0]^T$, $\mathbf{S}(t_2) = [0, 255]^T$, $\mathbf{S}(t_3) = [255, 0]^T$ and $\mathbf{S}(t_4) = [255, 255]^T$, the four “corner points” are observed. This should be the case for a pair of different contrasted images if they contains detail points. Of course, if parts (ii) and (iii) in the image are negligible, the edge of the scatter plots disappears and SWICA is likely to fail.

This explains why SWICA can match its first output \mathbf{Y}_1 to one of the sources (see Fig. 4.21.(d) and Fig. 4.21.(c)): the rotation of the whitened mixture scatter plot that yields the minimum value of $R(\mathbf{Y}_1)$ corresponds to Fig. 4.22.(d) But why does SWICA fail to recover the second source im-

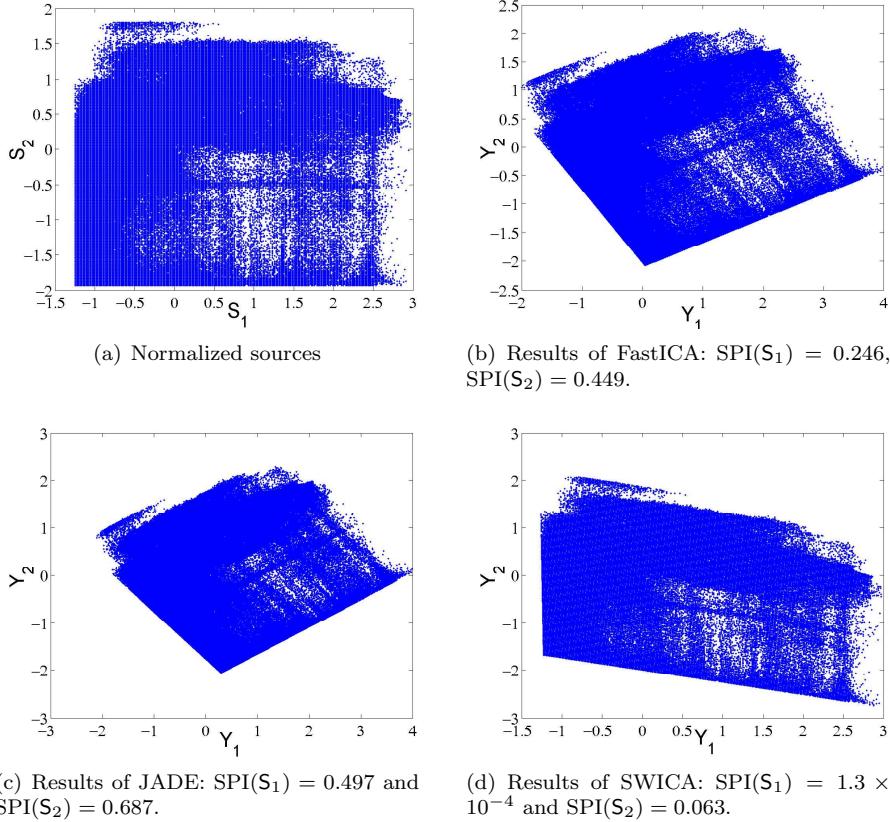


Figure 4.22. Scatter plots between normalized sources (a) and outputs of various ICA algorithms.

age? Actually, as for all ICA orthogonal contrasts, we whiten the mixtures beforehand and then constrain the demixing matrix \mathbf{B} to be orthogonal. Unfortunately, this constraint is too restrictive in our case and amounts to recovering sources that are not correlated. More precisely, on the one hand we know that $E[\mathbf{Y}\mathbf{Y}^T] = \mathbf{B}E[\mathbf{Z}\mathbf{Z}^T]\mathbf{B}^T = \mathbf{B}\mathbf{B}^T = \mathbf{I}_K$ because of the whiteness property (\mathbf{Z} denotes the random vectors of the whitened mixtures, to avoid confusion with a possible non-white mixture vector). On the other hand, we know that $E[\mathbf{S}\mathbf{S}^T]$ is not diagonal, which is contradictory.

- Extraction of the second source

The above-mentioned arguments explain why, when a deflation range-based BSS approach is used, one source can be recovered, whereas the other ones cannot be recovered when the source images are correlated and \mathbf{B} is constrained to be orthogonal.

In the easy case where $K = 2$, the second line of the \mathbf{B} matrix given by SWICA must be modified. Therefore, we minimize $R_N(\mathbf{Y}'_2) = R_N(\mathbf{b}'_2 \mathbf{Z})$ with respect to \mathbf{b}'_2 , where \mathbf{b}'_2 is not constrained to be orthogonal to \mathbf{b}_1 anymore. In order to avoid converging to \mathbf{Y}_1 , we take \mathbf{b}_2 (the second row of \mathbf{B} given by SWICA) as a first guess for \mathbf{b}'_2 . This procedure is applied only on the second output, without changing the first one, and allows separating the second source, as shown in Fig. 4.23. This procedure can be extended

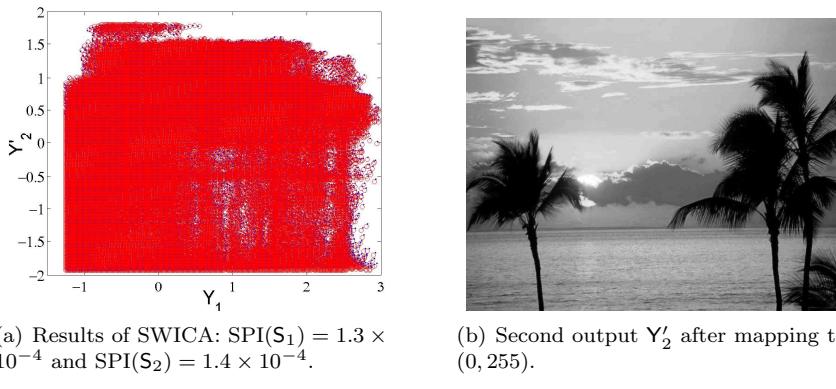


Figure 4.23. Results after extraction of the second source image: the low SPI values indicate that both sources are correctly recovered. Compared to Fig. 4.21.(f), the “white mountain” has been eliminated, such that the result better matches to the first source given in Fig. 4.21.(a)

for a larger number of source images, by deriving a deflation algorithm to correct the bias due to the orthogonality constraint. This is discussed in the following subsection.

4.4.1.2 NOSWICA: a non-orthogonal extension of SWICA

In the above method, a standard SWICA is applied and gives a first estimate of one of the sources; the second row is an orthogonal version of the first row, and thus because of the source correlation, a post-processing is performed on the second row of the demixing matrix in order to extract a second source. This two-step procedure is not optimal. Then, an extension of SWICA, called NOSWICA, involves a “smooth” orthogonal constraint; this non-rigid constraint is implemented as a penalization term preventing to extract several times a same source, without forcing a non-natural orthogonalization between the rows of \mathbf{B} . We shall first focus on a simultaneous separation scheme before considering a deflation method.

Simultaneous NOSWICA

Without prewhitening, recall that our goal is to find the smallest parallelepiped including the mixture joint density or, from the sample view point, to find the parallelepiped of smallest volume including all the sample points of the scatter

plot of the mixtures. As explained in Section 2.3.3, this can be found via the maximization of the criterion $\mathcal{C}_R(\mathbf{B})$ given by Eq. (2.41). The first term of this equation can be rewritten as [Hyvärinen et al., 2001, Pajunen, 1999]

$$\log |\det \mathbf{B}| = \sum_{i=1}^K \log \|(\mathbf{I}_K - \mathbf{P}_i)\mathbf{b}_i^T\| , \quad (4.42)$$

where $\mathbf{P}_i \doteq \mathbf{B}_i^T (\mathbf{B}_i \mathbf{B}_i^T)^{-1} \mathbf{B}_i$ is a projection matrix onto the subspace spanned by the columns of matrix $\mathbf{B}_i^T \doteq [\mathbf{b}_1^T, \dots, \mathbf{b}_{i-1}^T]$, that is the first $i-1$ rows of the demixing matrix \mathbf{B} . It can be checked that it is square, symmetric and idempotent. We take the natural convention $\mathbf{P}_1 = \mathbf{0}_K$.

Therefore,

$$\mathcal{C}_R(\mathbf{B}) = \sum_{i=1}^K \log \frac{\|(\mathbf{I}_K - \mathbf{P}_i)\mathbf{b}_i^T\|}{R(\mathbf{Y}_i)} . \quad (4.43)$$

This suggests that one actually minimizes the product of the output ranges subject to a “smooth orthogonality constraint”. We are looking for a vector \mathbf{b}_i such that i) the range of the associated output $\mathbf{Y}_i = \mathbf{b}_i \mathbf{X}$ is small and simultaneously ii) the distance between this vector and the hyperplane spanned by the previous rows of the demixing matrix is high.

Note that the numerator equals one under the orthonormality constraint of the rows of the separating matrix [Pajunen, 1999].

Deflation NOSWICA

One can trivially derive a deflation method based on the above simultaneous contrast by considering only a given row index i at a time. An i -th row is extracted by maximizing

$$\mathcal{C}_R(\mathbf{b}_i) = \log \frac{\|(\mathbf{I}_K - \mathbf{P}_i)\mathbf{b}_i^T\|}{R(\mathbf{Y}_i)} . \quad (4.44)$$

In particular, the recovering of the first row of \mathbf{B} is given by maximizing (the log of) $\frac{\|\mathbf{b}\|}{R(\mathbf{b}\mathbf{X})}$ with respect to \mathbf{b} . The “smooth orthogonality constraint” with respect to the previous rows of \mathbf{B} is implicitly contained in the projection matrix \mathbf{P}_i . This means that even in deflation schemes, the sources can be extracted sequentially without imposing a rigid orthogonality constraint.

In the previous range-based deflation approach presented in (2.44), the combination of pre-whitening step and orthogonality constraint implied that the algorithm necessarily led to recover a rectangle. Here, no pre-whitening is required even for the deflation approach, and thus the algorithm can yield, generally speaking, any non-singular parallelepiped. This is exactly what we need in image separation. Note however that a prewhitening step can be useful even if the dimensionality of the search space is not reduced in the next step. Whitening yields transformed outputs with similar variances; this will improve convergence of the NOSWICA algorithm when the source correlation is expected to be weak (see the discussion in Section 4.1.2).

$[\mathbf{BV}] = \text{NOSWICA}(\mathbf{X}(t))$

1. Whiten the mixtures using an eigenvalue value decomposition:
 - (a) remove the mean: $\mathbf{X}(t) \leftarrow \frac{1}{N} \sum_{t=1}^N \mathbf{X}(t)$
 - (b) compute the whitening \mathbf{V} matrix of $\mathbf{X}(t)$ by using Eq. (1.47)
 - (c) compute the projected whitened mixtures: $\mathbf{Z}(t) \leftarrow \mathbf{V}\mathbf{Z}(t)$
 2. Initialize $\mathbf{B} \leftarrow \mathbf{I}_K$.
 3. Extract the first source by minimizing the normalized source: $\mathbf{b}_1 \leftarrow \operatorname{argmin}_{\mathbf{b}} R(\mathbf{b}\mathbf{Z}/\|\mathbf{b}\|)$, and define $\mathbf{U} \leftarrow [\mathbf{b}_1^T]$.
 4. To extract the i -th source, with $1 < i \leq K$, do :
 - (a) compute the winning rows by minimizing the penalized range: $\mathbf{b}_i \leftarrow \operatorname{argmin}_{\mathbf{b}} R(\mathbf{b}\mathbf{Z}/\|\mathbf{b}^T - \mathbf{U}\mathbf{U}^T\mathbf{b}^T\|)$
 - (b) update \mathbf{U} : $\mathbf{U} \leftarrow [\mathbf{U}, (\mathbf{b}_i^T - \mathbf{U}\mathbf{U}^T\mathbf{b}_i^T)/\|\mathbf{b}_i^T - \mathbf{U}\mathbf{U}^T\mathbf{b}_i^T\|]$ to obtain a projection matrix onto the space spanned by the first $i-1$ rows of \mathbf{B} .
 5. return \mathbf{BV}
-

Table 4.3. Deflation NOSWICA algorithm.

It is intriguing and amusing to note that this smooth orthogonality constraint is exactly the same as the one that has been introduced in [Lee, Vrins, and Verleysen, 2006b] based on more geometrical aspects. This (deflation) “NOSWICA” algorithm is given in Table 4.3.

Observe that at each step of the algorithm of Table 4.3., the matrices $\mathbf{U}\mathbf{U}^T$ and \mathbf{P}_i (given below Eq. 4.43) are projection matrices onto the same subset. To see that, observe that the columns of \mathbf{U} form an orthonormal basis of the subspace spanned by the first $i-1$ rows of \mathbf{B} . The associated projection matrix is $\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$ which equals $\mathbf{U}\mathbf{U}^T$ as the columns of \mathbf{U} are orthonormal. Therefore, both $\mathbf{U}\mathbf{U}^T$ and \mathbf{P}_i are, at each step, projection matrices onto the subspace spanned by the first $i-1$ rows of the demixing matrix.

In practice, the exact (theoretical) range cannot be plugged in NOSWICA and has to be replaced by a sample-based estimator. In the sequel, it is always assumed that the m -averaged quasi-range is plugged in the method, just like in AVOSICA: $R(\mathbf{X}) \approx \langle R_N^{<m>}(\mathbf{X}) \rangle$. In the two following examples, the algorithm is applied on a toy example and on two image separation problems involving human faces.

Example 29 (Toy example) *It was explained in Section 2.3.4 that without a rigid orthogonalization constraint, the minimum range approach should be able to recover original sources provided that the edges of the source joint density resemble the edges of independent source densities, that is the source density convex hull is a rectangle. The correlation “inside” the source density should have no impact on the result. To this end, dependent source signals are built as*

$S_i(t) = C_i(t)$ with probability α and $S_i(t) = P_i(t)$ with probability $1 - \alpha$. The $P_i(t)$ are mutually independent random variables while the $C_i(t)$ are mutually correlated.

Setting $K = 2$ and $\alpha = 0.5$, each source is an equiprobable mixture of samples of both correlated and fully independent variables. The independent part $P_i(t)$ of the sources is built from samples drawn from the uniform density in the interval $(-1, +1)$. The dependent part of the sources $C_i(t)$ is built as follows. First, two independent zero-mean unit-variance Gaussian densities are sampled. Next, the obtained vectors are mixed by pre-multiplying them with the matrix $[0.2, 0.4; 0.4; 0.2]$, in order to correlate their components. Finally, all values greater than one in absolute value are replaced by ± 1 . The covariance matrix of the resulting sources is:

$$\Sigma_S = \begin{bmatrix} 0.25 & 0.07 \\ 0.07 & 0.25 \end{bmatrix} .$$

The first plot in Fig. 4.24. shows a 500 points source scatter plot obtained according to the above building scheme. The second plot displays mixtures of these sources. The third plot shows the whitened mixtures. The three last plots illustrate the results of FastICA (deflation, pow3), AVOSICA and NOSWICA. Only NOSWICA yields the expected result.

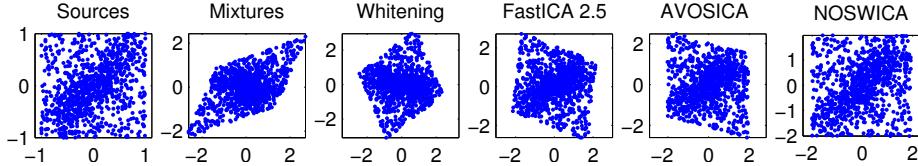


Figure 4.24. Example 29: computer-generated composite densities. The six plots show respectively the two sources, two random mixtures, the whitened mixtures and results of FastICA, AVOSICA and NOSWICA ($N = 500$, $m = m^\sharp(N) = 12$).

Example 30 (Mixtures of real images I) For this experiment, color pictures of human faces were cropped in order to share the same size (192 by 144 pixels) and converted to grayscale using `rgb2gray` in Matlab, as shown in Fig. 4.25.

Next, the rows of pixels of each image are concatenated to obtain three row vectors that contains the observations of each source. Finally, those vectors are standardized so that sources have zero mean and unit variance. The leftmost plots in Fig. 4.25. show the histograms of the sources. Computing the covariance matrix of these sources leads to

$$\Sigma_S = \begin{bmatrix} 1.00 & -0.27 & -0.25 \\ -0.27 & 1.00 & 0.51 \\ -0.25 & 0.51 & 1.00 \end{bmatrix} . \quad (4.45)$$

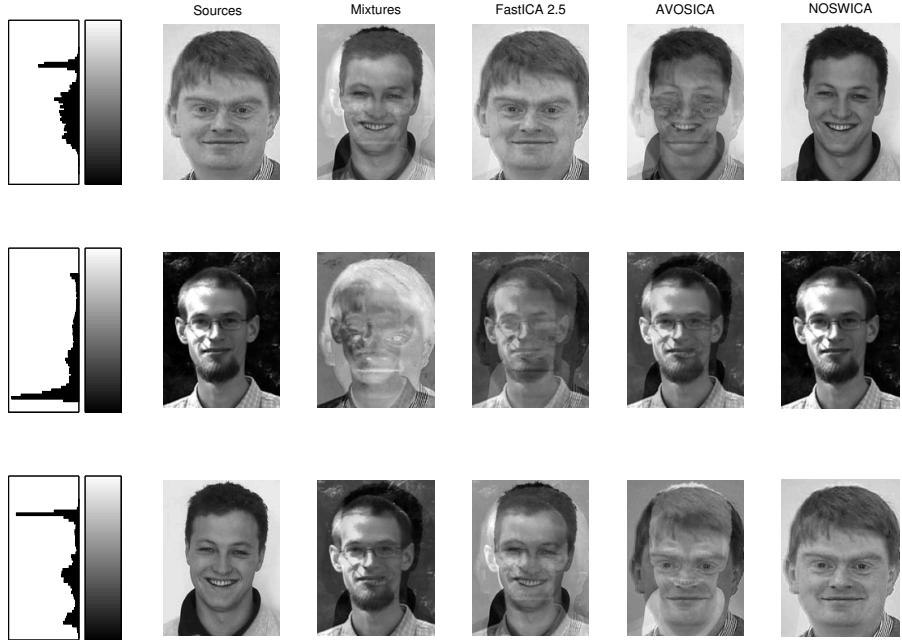


Figure 4.25. Example 30: mixtures of real pictures. The two leftmost columns shows the sources (from top to bottom: Pr. M. Verleysen, Dr J.A. Lee and the present author) and their histogram. The third column consists of three random mixtures of the sources (mapped to [0, 255] for readability purposes). The three rightmost columns display results of FastICA, AVOSICA and NOSWICA ($N = 27648$, $m = m^\sharp(N) = 224$).

Pictures corresponding to three random linear combinations of the sources are shown in the third column of Fig. 4.25. The three rightmost columns show the results of FastICA (version 2.5, deflation, $\text{pow}3$), AVOSICA ($m = m^\sharp(27648)$) and NOSWICA ($m = m^\sharp(27648)$). As in the case of the toy example, NOSWICA clearly outperforms the two other algorithms. The quality of the results may be assessed using the SPI performance index (Eq. (4.41)).

In Fig. 4.26., the three algorithms are ran 100 times with different mixtures and histograms of the SPIs are displayed for each source or for all of them (average SPI); we have chosen $m = 500 \approx 2m^\sharp(N)$ due to the very large number N of samples; this leads to a slightly better result than for $m = m^\sharp(N)$. As it can also be seen in Table 4.4., the average result of NOSWICA is pretty good. On the other hand, its robustness is more disappointing: the variance of the SPIs is not negligible. Because the orthogonality constraint is relaxed in NOSWICA, the space of solutions is larger than for the two other algorithms. This could account for the result variability.

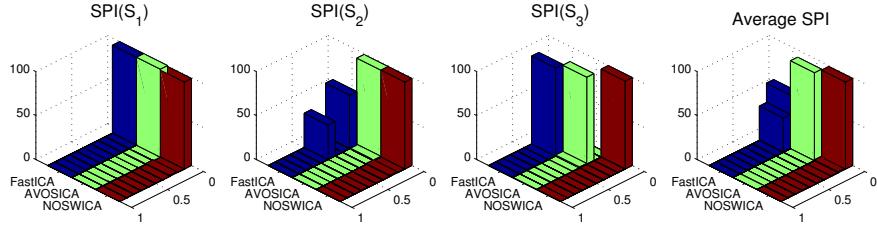


Figure 4.26. Histograms of the SPIs ($N = 27648$, $m = 500$); the sources are the left-most pictures of Figure 4.25., “top-down” ordered. Lowest values are the best ones.

SPI	FastICA 2.5	AVOSICA	NOSWICA
Source 1	0.01 (5×10^{-6})	9×10^{-4} (8×10^{-8})	5×10^{-3} (7×10^{-4})
Source 2	0.24 (3×10^{-2})	0.07 (2×10^{-5})	6×10^{-3} (5×10^{-4})
Source 3	0.31 (2×10^{-5})	0.27 (7×10^{-5})	6×10^{-3} (3×10^{-5})
Average	0.19 (3×10^{-3})	0.11 (1×10^{-5})	6×10^{-3} (2×10^{-4})

Table 4.4. Mean and variance (between parentheses) of the SPIs for the example in Fig. 4.25., over 100 trials with different random mixtures, $m = 500$. Best values are bolded; the sources are the left-most pictures of Figure 4.25., “top-down” ordered.

Example 31 (Mixtures of real images II) Figure 4.27. shows the same experiment as in Example 30 but in which the source pictures are those of the thesis Jury (PhD advisor excluded; he was already involved – 101 times! – in the previous experiment). The source covariance matrix is in this case:

$$\Sigma_S = \begin{bmatrix} 1 & -.33 & -.04 & .55 & .06 \\ -.33 & 1 & .20 & -.33 & -.06 \\ -.04 & .20 & 1 & -.11 & -.05 \\ .55 & -.33 & -.11 & 1 & .08 \\ .06 & -.06 & -.05 & .08 & 1 \end{bmatrix}.$$

In spite of the fact that some of the source images are “pixelized” (their common size is 74×77 pixels only), one can see in the figure that the results are less good. However, note that AVOSICA and NOSWICA still outperform FastICA⁵. A first reason for explaining that observation results from the fact that the pictures are not saturated enough. Indeed, in order that range-based ICA techniques perform, one needs to have sample points in the corner of the scatter plot; in other words,

⁵A single experiment is shown here as it is given to emphasize the limits of the method only. Furthermore, the convergence problems encountered by FastICA do not allow us to compare the results easily.

we need “sharp-boundaries histogram figures”. This is not the case for several pictures (e.g. C. Jutten and P. Lambert). The picture of P. Delsarte is weakly contrasted while, on the contrary, those of E. Oja and J.-D. Legat match our requirements. We mention that the most “suitable pictures” (in the sense of NOSWICA) are often pretty well extracted. FastICA almost fails to recover more than one source image (we mention in passing that in many cases, Erkki Oja – one of the inventors of the algorithm – is surprisingly correctly extracted by this algorithm). But this also emphasizes other problems and, in particular, the effect of a too weak ratio of sample points-to-source dimensionality. Compared to Example 30, the number of samples has decreased (from 27648 to 5698), while the source dimensionality is higher (5 instead of 3). When the number of sources K increases, the theoretical number of corners points (2^K) increases exponentially. But the source space becomes more and more empty (to see that, imagine N points in a unit square, a unit cube, ...); this is often referred to as “the empty space phenomenon” or the “curse of dimensionality”. Therefore the probability to observe such “corner points” decreases if the density of the sources have flat tails (the joint density is the product of the marginals as the sources are i.i.d.). The common effect of too weakly contrasted sources (flat tails) and their number (i.e. the number of corner points) is that the probability to observe “interesting points”, that would lead range-based approaches to the right solution, decreases. Observe however that range-based approaches become robust to the dimensionality of the source space when the densities are uniform or “V”-shape because in that case, more and more points are observed in the boundary (this robustness will be emphasized on a variant of the NOSWICA algorithm in Section 4.4.2.4). Figure 4.28. emphasizes these origins of the problem. In this experiment, we have artificially saturated the pictures of the Jury members (sorry about that): all the pixels having a value less than 45 are set to 0 and those higher than 215 are set to 255. This leads to much more contrasted pictures. As there are more points near 0 and near 255 (see the related histograms), there is more chance that corner points (like [0, 0, 0, 0, 0], [255, 0, 0, 0, 0], ...) will be observed. This would clearly help the range-based algorithms to succeed.

4.4.2 Application to lower- or upper-bounded sources with possible infinite range

In this section, the application field of minimum range methods is extended from bounded sources (with finite range) to sources with possibly infinite range, but having either a lower or an upper bound. This extension was proposed in [Lee, Vrins, and Verleysen, 2006a], in the framework of MLSP 2006 competition on data analysis.

4.4.2.1 LABICA

A variant of the minimum-range approach is proposed for extracting source signals that are possibly bounded on one side only: $\min(\sup(S_i), |\inf(S_i)|) < \infty$



Figure 4.27. Example 31: mixtures of real pictures. The two leftmost columns shows the sources (from top to bottom: Pr. P. Delsarte, Pr. P. Lambert, Pr. C. Jutten, Pr. E. Oja and Pr. J.-D. Legat) and their histogram. The third column consists of three random mixtures of the sources (mapped to [0, 255] for readability purposes). The three rightmost columns display results of FastICA, AVOSICA and NOSWICA ($N = 5698$, $m = m^\sharp(N) = 77$).

while the minimum range approach required that a stronger condition holds: $\max(\sup(S_i), |\inf(S_i)|) < \infty$. As sources can only be recovered up to a sign change, this leaves two solutions: one can either maximize the infimum of the whitened mixtures or minimize the supremum, which can be related to Erdogan's approach [Erdogan, 2006] (minimization of the supremum for symmetric bounded signals). These two possibilities can be merged into a single objective function, namely the Least Absolute Bound (LAB):

$$\text{LAB}(\mathbf{b}_i Z) \doteq \min\{-\inf(\mathbf{b}_i Z), \sup(\mathbf{b}_i Z)\} , \quad (4.46)$$

which has to be minimized: $\mathbf{b}_i^* \doteq \min_{\mathbf{b}_i} \text{LAB}(\mathbf{b}_i Z)$. Defining $\mathcal{C}_{\text{LAB}}(\mathbf{b}_i) = -\text{LAB}(\mathbf{b}_i Z)$, we are led to maximize $\mathcal{C}_{\text{LAB}}(\mathbf{b}_i)$.

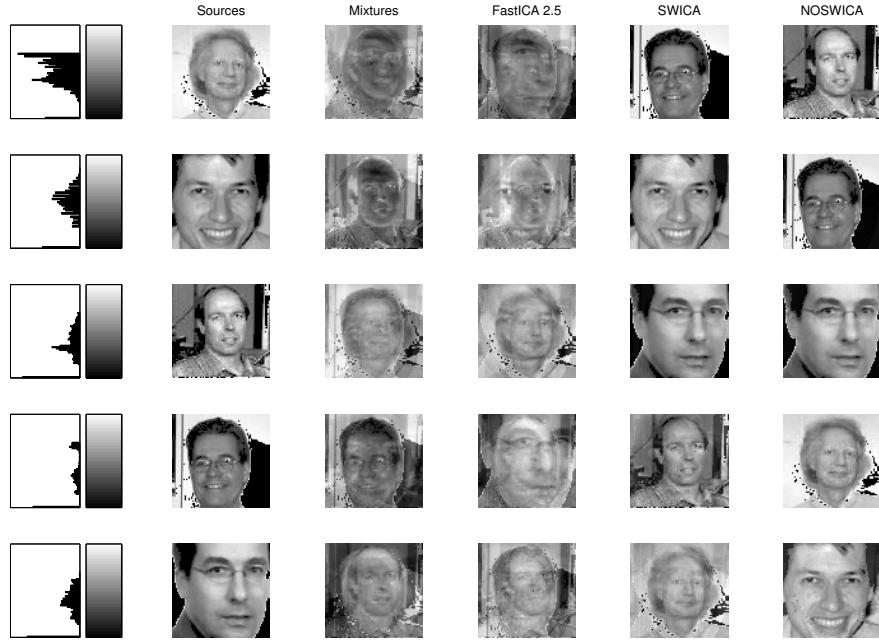


Figure 4.28. Example 31: mixtures of (artificially) saturated pictures. The two leftmost columns shows the artificially saturated sources and their histogram. The third column consists of three random mixtures of the sources (mapped to [0, 255] for readability purposes). The three rightmost columns display results of FastICA, AVOSICA and NOSWICA ($N = 5698$, $m = m^\sharp(N) = 77$).

The above function assumes working on whitened mixtures \mathbf{Z} . Otherwise, a normalization with respect to the output variance or demixing vector \mathbf{b}_i is needed, just as for the range.

Under the whiteness assumption on the independent sources, the minimization of the LAB criterion is equivalent to the maximization of

$$\tilde{\mathcal{C}}_{\text{LAB}}(\mathbf{w}_i) \doteq -\min\{-\inf(\mathbf{w}_i \mathbf{S}), \sup(\mathbf{w}_i \mathbf{S})\} \quad (4.47)$$

with respect to \mathbf{w}_i and $\mathbf{w}_i^* = \mathbf{b}_i^* \mathbf{V} \mathbf{A}$ where $\mathbf{w}_i^* \doteq \max_{\mathbf{w}_i} \tilde{\mathcal{C}}_{\text{LAB}}(\mathbf{w}_i)$ since $\tilde{\mathcal{C}}_{\text{LAB}}(\mathbf{b}_i \mathbf{V} \mathbf{A}) = \mathcal{C}_{\text{LAB}}(\mathbf{b}_i)$ according to our notation convention.

As the LAB only focuses on the bounded part of the random variable, it behaves like the range from the local minima point of view. To see that (in the $K = 2$ case), compare the angular variation of $R(\mathbf{Y}_1)$ and $\text{LAB}(\mathbf{Y}_1)$ as a function of the mixing angle $\theta = \phi + \varphi$ in Fig. 4.29.; the two sources are assumed to be double-bounded (left panel) or lower-bounded (right-panel). When the sources

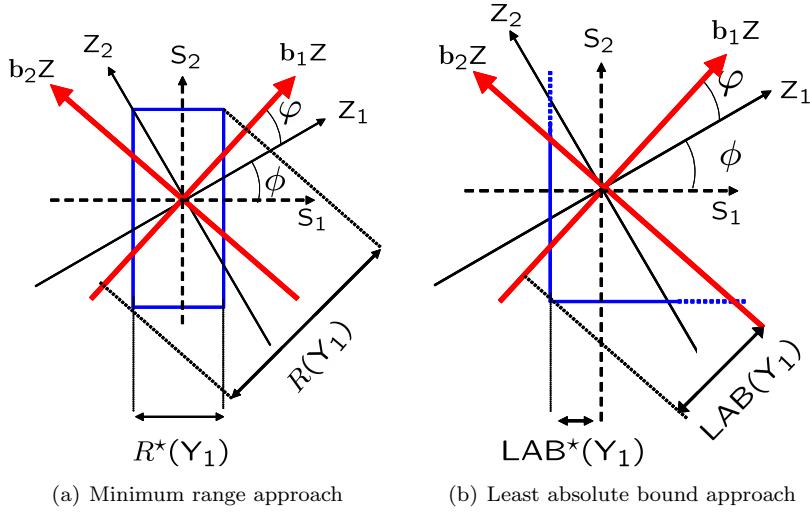


Figure 4.29. Graphical interpretation of the minimum range and least absolute bound approaches; the minimum values of the criteria are labelled ‘*’ and are seen to correspond to the extraction of one particular source. The solid lines represents the support boundaries and the large solid arrows represent the current axes.

are lower- and upper-bounded, both criteria are similar. The same holds true when at least one of the sources is double bounded; in particular, $\text{LAB}(\mathbf{Y}_1)$ is finite, whatever the value of the transfer angle θ . Things are different if the two sources have an infinite extremum (assumed to be the supremum in Fig. 4.29.(b)). For $\theta \in [0, \pi/2] \cup [\pi, 3\pi/2]$, the LAB behaves like the range, and on the other quadrants (boundaries excluded), it may be (theoretically) infinite. In practice, the LAB criterion will take very large values in the second and fourth quadrants, so that the global minimum must be in the quadrant where the LAB behaves similarly as the range. Consequently, the location of the local minimum point is still $\theta \in \{k\pi/2 | k \in \mathbb{Z}\}$, and the global minimum corresponds to the extraction of the source having the lowest absolute bound.

4.4.2.2 Practical estimation

Estimating a finite extreme point of a distribution is not an easy task but is close from range estimation. Therefore, the order-statistics based method that has been presented in Section 4.2 can be used. It is proved to yield promising results as shown in the following subsection.

Assuming that \mathbf{Y}'_i denotes the sorted i -th output, the contrast estimator can be written as

$$\hat{\mathcal{C}}_{\text{LAB}}(\mathbf{b}_i) \doteq - \min \left\{ -\frac{1}{m} \sum_{k=1}^m \mathbf{Y}'_i(k), \frac{1}{m} \sum_{k=1}^m \mathbf{Y}'_i(N+1-k) \right\}, \quad (4.48)$$

where $\mathbf{Y}'_i(k)$ is the k -th lowest value of $\mathbf{Y}_i(t)$, $\mathbf{Y}'_i(N+1-k)$ the k -th highest one and m is an integer between 1 and $\lfloor N/2 \rfloor$. We shall use a similar “hat” notation for the empirical counter part of $\text{LAB}(\mathbf{b}_i \mathbf{Z})$. The sample size N is supposed to be large enough so that the accuracy of the above estimator is sufficient. In other words, the observed minimum sample should be close to the true theoretical infimum (and likewise for the observed maximum and theoretical supremum) to ensure that the empirical criterion will behave similarly as the theoretical one. Under the same condition, $\frac{1}{m} \sum_{k=1}^m \mathbf{Y}'_i(k) \simeq \min_t \mathbf{Y}_i(t)$, and likewise for $\frac{1}{m} \sum_{k=1}^m \mathbf{Y}'_i(N+1-k)$.

In the noise-free case, m can be taken close to one; otherwise, m must be slightly increased. In a mimetic manner, we can take $m = m^\sharp(N)$.

4.4.2.3 Optimization scheme for “hard” ICA problems

In this section we shall give an ICA algorithm that aims at solving “hard” ICA problems (mainly: source separation from ill-conditioned and large-scale mixtures).

The optimization strategy is basically the same as the one of AVOSICA, except that several additional precautions have been taken:

- The prewhitening step involved SVD rather than the usual EVD to save robustness;
- the possibly (say p) last signals that have been poorly whitened are simply discarded from the mixture set; only $K' \leq K$ ($K' = K - p$) are considered in what follows and thus K' sources shall be estimated; the last p source signals are simply guessed by setting them equal to the p badly decorrelated signals. Hence, K' sources can be recovered in a satisfactory way because “unreliable signals” have been discarded, but p of them (with hopefully small p) are really bad estimates;
- Rigid orthogonalization is avoided and the demixing matrix is only prevented to be singular (the minimum allowed angle between two of its rows is of $\pi/12$);
- the computation cost should be minimized as we are dealing with large-scale problems. Hence the number of criterion evaluations has to be kept quite small.
- An alternative to gradient-ascent has been recently proposed [Lee, Vrins, and Verleysen, 2006a]. In this paper, it is proposed to adapt the convergence rate of the algorithm to “how close we are from a point far from the solution”. Such a point is clearly the “corner” of the scatter plot.

Let us briefly detail the last item. Let \mathbf{u} denote the direction of the average of the m points used in the criterion evaluation before projection on the vector \mathbf{w} . These points are easily found because they are the z -points corresponding

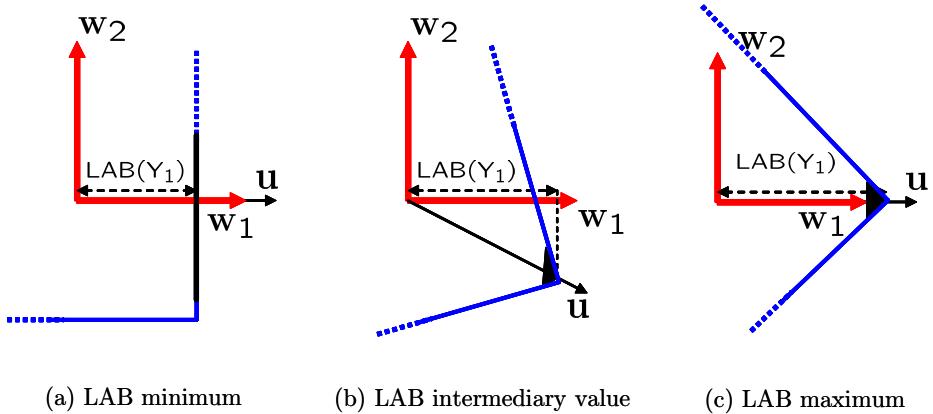


Figure 4.30. Principle of the proposed algorithm in the 2D case. In those three examples, vector \mathbf{u} points towards sample points (symbolized by the bold black line or black triangle) that determine the value of $\hat{\text{LAB}}(\mathbf{w}_S)$. This provides information about how to update \mathbf{b}_1 : if $\mathbf{u} \neq \mathbf{b}_1$, then rotate \mathbf{b}_1 away from \mathbf{u} (see text).

to the m output samples involved in the “winner minimum” in Eq. (4.48). Most of the time, these points are grouped around a corner (this becomes more and more true for increasing N), but when we are close to a solution, these points can be spread along a face of the scatter plot (the edge when $K = 2$) which is perpendicular to the vector \mathbf{w} ; see Figure 4.30. The interesting point is that the criterion reaches a local extremum when \mathbf{w} is co-linear with \mathbf{u} . Depending on the case, this local extremum may be a minimum (Fig. 4.30. (a).) or a maximum (Fig. 4.30. (c).) point.

The cosine between the current vector \mathbf{w} and the closest corner \mathbf{u} is an element of how close \mathbf{w} is from a point that is *far* from being a good solution. Except at the solution point, if the last cosine is large, \mathbf{w} can be largely updated but on the contrary, if the angle between these vectors is large, the update should be moderate. In other words, the projection of \mathbf{u} on a plane orthogonal to \mathbf{w} can be seen as a “pseudo-gradient”, which acts as a gradient (Fig. 4.30.). This projection is zero when we are at an extremum point (local maximum when \mathbf{w} points to a corner, local minimum when \mathbf{w} is perpendicular to an edge of the density). Note that between two iterations, \mathbf{u} can “jump” between two directions because the closest (i.e. winner) corner may differ, even with a small learning rate, when we are close to the solution. Another drawback is that like the gradient, it does not decrease smoothly to zero when approaching the solution; both can take large values right near a solution point. However, this should not prevent convergence if the learning rate decreases. Indeed, the optimization scheme will approach the solution point, when \mathbf{w} tends to be perpendicular to an edge. When we are very close to the solution, \mathbf{u} is nearly perpendicular to the edge, and thus nearly co-linear with \mathbf{w} ; the pseudo-gradient indicates that

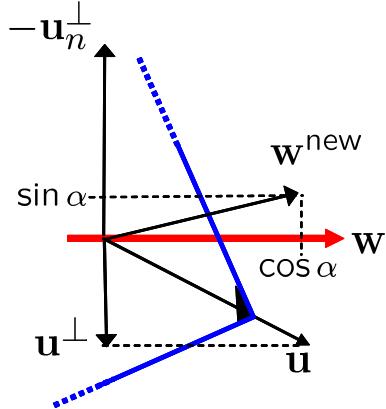


Figure 4.31. Illustration of step 4.(e) in the algorithm of Table 4.5.

a large update is needed (because we cannot make the difference with the case of Fig. 4.30. (c).). However, in that situation, the learning rate should be small (and in any case, will decrease), so that in practice, the update is kept “quite small”. Observe that similar things could be done by using the gradient even if the gradient does not exist at the solution point (just like the closest corner point does not really have a meaning when we are at the solution point). However, the corner-based approach has a nice advantage: the computation of \mathbf{u} is trivially obtained through the contrast evaluation step.

The detailed algorithm is proposed in pseudo-code in Table 4.5. The main idea (step 4.(e)) is to update \mathbf{b} according to \mathbf{u} : if $\mathbf{b} \neq \mathbf{u}$, we shall move \mathbf{b} away from \mathbf{u} .

Note that in the above algorithm, α is a learning rate, even if used via circular functions. The idea behind the step 4.(e) is illustrated in Fig. 4.31.

Remark 23 In order to avoid the cumulation of errors, which may have critical consequences when K is large, it has been explained that a smooth orthogonality condition might be better than a rigid one. There are two possibilities to implement this non-rigid orthogonality constraint. In the first one (named “A”), we look for a non-orthogonal demixing matrix in $R^{K \times K}$ based on a penalized criterion. A second possible way to deal with a non-rigid orthogonalization (named “B”) we apply a two-step procedure to extract an i -th source. This procedure consists in 1) arbitrarily setting the i -th row of \mathbf{B} such that it is not too close (from an angular point of view) to the previous $i-1$ rows that already correspond to the extraction of $i-1$ distinct sources and 2) based on this first guess, to relax the row-orthogonality constraint by minimizing $LAB(\mathbf{Y}_i)$ (the rows of \mathbf{B} are thus, at the end, only constrained to have a unit-norm, just as in A)). Note that in B), there is a “re-initialization” procedure if \mathbf{b}_i converges to one of the first $i-1$ rows of the demixing matrix. The approach A) was sketched in Figure 4.3., but B) was used in the algorithm of Table 4.5. This is because based on the simula-

$[\mathbf{B}, \mathbf{V}] = \text{LABICA}(\mathbf{X}(t))$

1. Whiten the mixtures using a singular value decomposition:
 - (a) Center the sample by removing its mean: $\mathbf{X}(t) \leftarrow \mathbf{X}(t) - \frac{1}{N} \sum_{t=1}^N \mathbf{X}(t)$.
 - (b) Compute the SVD of the centered sample: $\mathbf{X}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$.
 - (c) Compute $\mathbf{Z}(t)$ directly: $\mathbf{Z} = \sqrt{N} \mathbf{U}^T$.
(Depending on the convention, \mathbf{U} is either $N \times K$ or $N \times N$; in the latter case, keep only the K first columns of \mathbf{U} .)
 2. Discard the p incorrectly whitened mixtures (i.e. rows $\mathbf{Z}_i(t)$ having a variance lower than one and/or nonzero covariances).
 3. Compute the radial projection of \mathbf{Z} on the unit sphere:

$$\mathbf{Z}^\circ(t) = \frac{\mathbf{Z}(t)}{\|\mathbf{Z}(t)\|} \text{ for } 1 \leq t \leq N.$$
 4. To extract the i -th source, with $1 \leq i \leq K - p$, do:
 - (a) Initialize \mathbf{b}_i to any random direction and the update angle α to $\pi/4$.
 - (b) Check loose orthogonality: if for some $j < i$ the inequality $|\mathbf{b}_i \mathbf{b}_j^T| < \cos(\pi/12)$ holds then make \mathbf{b}_i orthogonal to all \mathbf{b}_j : $\mathbf{b}_i \leftarrow \mathbf{b}_i - \sum_j \mathbf{b}_i \mathbf{b}_j^T \mathbf{b}_j$; $\mathbf{b}_i \leftarrow \frac{\mathbf{b}_i}{\|\mathbf{b}_i\|}$.
 - (c) Compute i -th ICA output: $\mathbf{Y}_i(t) = \mathbf{b}_i \mathbf{Z}(t)$ for $1 \leq t \leq N$.
 - (d) Estimate the LAB of $\mathbf{Y}_i(t)$ using mean order statistics:
 - Determine the indexes of the m lowest and m highest values of $\mathbf{Y}_i(t)$.
 - Average the two corresponding sets of values to obtain the infimum and supremum of $\mathbf{Y}_i(t)$; keep their minimum absolute value as in (4.48) to obtain $\hat{\text{LAB}}(\mathbf{Y}_i)$.
 - Use the same indexes to compute the direction \mathbf{u} as the average of the corresponding columns of \mathbf{Z}° .
 - If $\mathbf{u} \neq \mathbf{b}_i$, make \mathbf{u} orthogonal to \mathbf{b}_i and normalize it: $\mathbf{u}^\perp = \mathbf{u} - \mathbf{u} \mathbf{b}_i^T \mathbf{b}_i$; $\mathbf{u}_n^\perp = \frac{\mathbf{u}^\perp}{\|\mathbf{u}^\perp\|}$, $\mathbf{u} \leftarrow \mathbf{u}_n^\perp$.
 - (e) Update \mathbf{b}_i and α :
 - Compute $\mathbf{b}'_i = \cos(\alpha) \mathbf{b}_i - \sin(\alpha) \mathbf{u}$ and $\hat{\text{LAB}}(\mathbf{b}'_i \mathbf{Z})$ (see step (d) above).
 - If $\hat{\text{LAB}}(\mathbf{b}'_i \mathbf{Z}) < \hat{\text{LAB}}(\mathbf{Y}_i)$, then let $\alpha \leftarrow 1.01\alpha$ and $\mathbf{b}_i \leftarrow \mathbf{b}'_i$, else $\alpha \leftarrow \alpha/1.2$.
 - (f) Go back to step 4(b) if convergence is not attained.
 5. Append the p incorrectly whitened mixtures to the extracted sources: $\forall i > K - p$, $\mathbf{Y}_i(t) \leftarrow \mathbf{Z}_i(t)$.
-

Table 4.5. LABICA: ad hoc deflation procedure to minimize $\hat{\text{LAB}}$. After robust SVD-based whitening, sources are extracted one-by-one, with a loose orthogonality constraint preventing error accumulation. The gradient is replaced by contrast-dependent information: the closest support corner direction.

tion results shown below, the procedure B) is more efficient than A). We are not able to find clear arguments for justifying the results of the different approach; it seems that in practice, in a high-dimensional source space, the penalization jeopardizes the recovering of the sources so that it is preferable to focus on the criterion only.

4.4.2.4 Application of LABICA to the MLSP'06 competition benchmark: performances analysis

LABICA, the extension of SWICA to sources that are bounded on one side only, has been tested on a competition benchmark organized in the framework of the *2006 IEEE Workshop on Machine Learning for Signal Processing*. There were four subproblems to deal with, each of them containing an equiprobable mixture of source signals drawn from two densities. The first source density was the uniform density with support set defined in $[0, 1]$ (double-bounded sources), the second is a kind of sparse density taking values in \mathbb{R}^+ (lower-bounded sources). For more details, see the data analysis competition announcement in the Appendix A.

In order to assess the quality of the source recovery, the competition resorts to the Signal-to-Interference Ratio (SIR), which involves the transfer matrix $\mathbf{W} = \mathbf{BA}$ and can be defined as follows:

$$\begin{aligned} \text{SIR} &= 10 \log_{10} \text{SPI} \\ &= \frac{1}{K} \sum_{i=1}^K 10 \log_{10} \frac{\max_j W_{ij}^2}{\sum_{j=1}^K W_{ij}^2 - \max_j W_{ij}^2}, \end{aligned} \quad (4.49)$$

and expressed in dB. Within the framework of the competition, the SIR is used in a Monte-Carlo process. The SIR should be higher than 15dB for at least 90% of the runs, i.e. $P_{90} > 15\text{dB}$, where P_{90} is the 90-th percentile of the SIR. Basically, this means that the SIR should be higher than 15dB with a probability of 90%.

The algorithm performances were analyzed under various angles; four subproblems were proposed to test the robustness to high dimensional source space, small sample set, mixing matrix with low condition number (defined as the ratio of its largest singular value to the smallest one) and to additive white Gaussian noise. The behavior of the algorithm was tested in the four cases by finding the limit conditions such that the success criterion $P_{90} > 15\text{dB}$ holds. More explicitly, we had to find

1. *the largest number K of sources* with fixed sample size ($N = 5000$) and random mixing matrix,
2. *the smallest number N of samples* with fixed number of mixtures ($K = 50$) and random mixing matrix,
3. *the largest number K of sources* with fixed sample size ($N = 5000$) where \mathbf{A} is a Hilbert matrix multiplied by a random Givens matrix (a Hilbert matrix

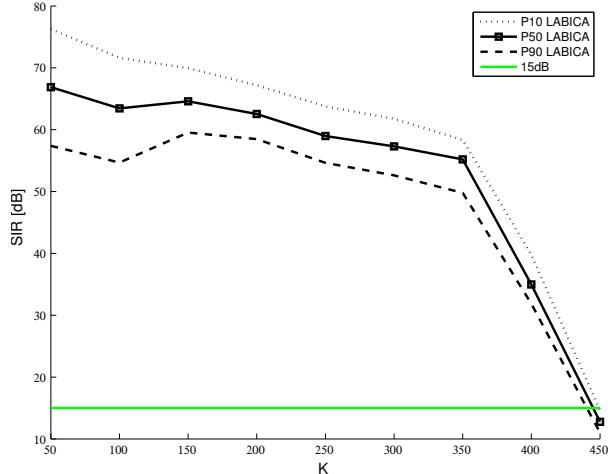


Figure 4.32. Results for Subproblem 1: SIR performances vs number of sources K for 20 Monte-Carlo runs of LABICA and FastICA with 5000 sample points. $P_{90} > 15\text{dB}$ holds for more than 300 sources.

$\mathbf{H} \in \mathbb{R}^{K \times K}$ is defined element-wise as $H_{ij} \doteq (i + j - 1)^{-1}$, and becomes more and more difficult to invert for increasing K as the determinant tends rapidly to zero),

4. *the largest noise variance* with $K = 50$, and $N = 1000$,

under the condition that the success criterion is met.

- *Subproblem 1: Large-scale problem.* In this first subproblem, the sample size is fixed ($N = 5000$) and the number of sources/mixtures is growing ($K > 50$). The algorithm proposed in Table 4.5. solves it for a quite large number of mixtures. Graphical results in Fig. 4.32. show that outstanding SIR values are attained for more than 400 mixtures (P_{90} is still higher than 30dB). Processing so many mixtures obviously requires long computation time, even with the fastest algorithms (e.g. FastICA), and justifies the restriction to only 20 Monte-Carlo runs.
- *Subproblem 2: Small training set problem.* In this second problem, the number of sources is kept constant ($K = 50$) but the sample size N varies. The results are shown in Fig. 4.33. for two algorithms: the proposed one and FastICA (with ‘gaus’ nonlinearity and fine tuning enabled). As can be seen, less than 250 sample points are required to achieve a SIR greater than 15dB in 90% of the cases.
- *Subproblem 3: Highly ill-conditioned problem.* In this third subproblem, the mixing matrix is the product of a Hilbert matrix with a random Givens

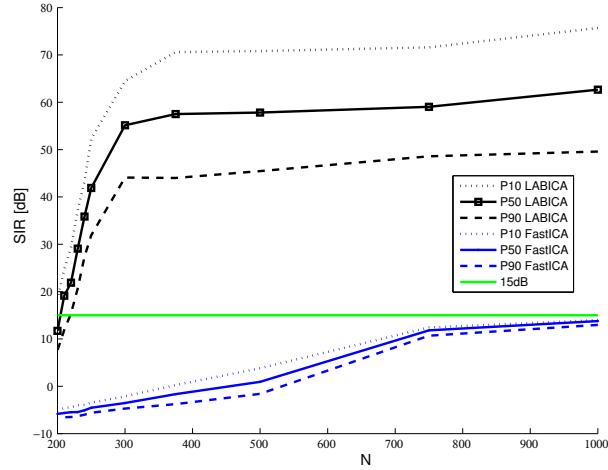


Figure 4.33. Results for Subproblem 2: SIR performances vs sample size N for 100 Monte-Carlo runs of LABICA and FastICA with 50 sources. Less than 250 observations are needed to achieve $P_{90} > 15\text{dB}$. All the P_X curves of FastICA are below the 15dB threshold line.

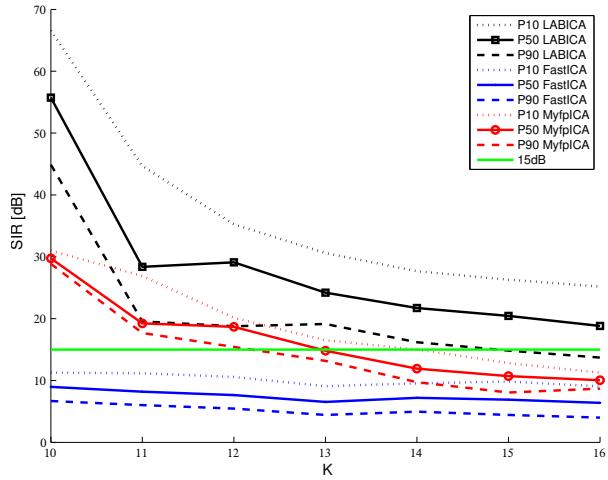


Figure 4.34. Results for Subproblem 3: SIR performances vs the number of sources K for 100 Monte-Carlo runs of LABICA ($m = N/200$), FastICA and MyfpICA with 5000 sample points. $P_{90} > 15\text{dB}$ holds up to 14 sources.

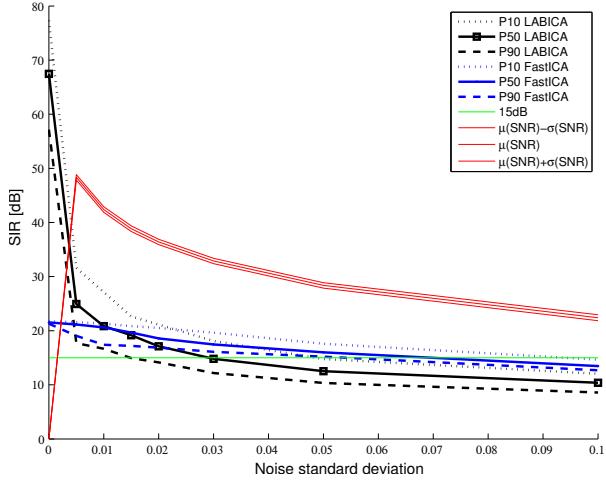


Figure 4.35. Results for Subproblem 4: SIR performances vs the noise standard deviation, for 100 Monte-Carlo runs of LABICA and FastICA with 5000 sample points and 50 sources; the corresponding SNR curve is plotted alongside.

matrix. Hence, as the number of mixtures is growing, the separation problem gets more and more ill-conditioned. The results are shown in Fig. 4.34. for three algorithms: the proposed one, FastICA (as above) and a ‘hacked’ version of FastICA. The latter, called MyfpICA, works with a SVD-based whitening stage and a kurtosis-driven nonlinearity (either ‘kurt’ or ‘gaus’ depending on the kurtosis). In this subproblem, achieving a correct whitening is the main difficulty. The proposed algorithm brings a significant performance gain by using the SVD of the centered sample instead of the EVD of the sample covariance matrix. However, beyond 10 mixtures in this problem, the determinant of the mixing matrix \mathbf{A} is so close to zero that no more than 10 mixtures can be whitened properly, even with the SVD (to understand why, note that with only $K = 6$, $\det \mathbf{H} \approx 1E-18$). It has been experimentally observed that additional mixtures after whitening are actually not white; some of them may be correlated and/or have a variance lower than one. In this situation, the trick consists in temporarily discarding these still correlated mixtures after whitening, as proposed in Section 4.4.2.3, so that the separation algorithm can run in good conditions.

- *Subproblem 4: Noisy mixtures problem.* The value of estimator $\hat{\mathcal{C}}_{\text{LAB}}(\mathbf{b}_i)$ relies on a few sample points only, namely on m sample points with $m \ll N$. Consequently the proposed approach is expected not to be very robust against noise and outliers, especially with low values of m . As can be seen in Fig. 4.35., the quality of the results is rapidly decreasing as the noise variance is growing.

4.5 CONCLUSION OF THE CHAPTER

In this chapter, we have proposed order-statistic-based range estimators for minimum range BSS approaches. Within the simultaneous ICA framework, Pham was the first to propose to use order-statistics and, more specifically, to use them to approximate the range [Pham, 2000]. However, order statistics can also be used in ICA methods in a different manner, as e.g. to estimate the source pdf/cdf or score functions. For more details about these approaches, we refer to [Even, 2003, Blanco and Zazo, 2004] and to the references therein. Some properties, advantages and drawbacks of our range estimator have been discussed, both from the range estimation and separation performance viewpoint. An important issue was to develop an ad-hoc optimization procedure for our range-based criterion because its gradient does not exist on the sought (solution) points. A simple algorithm based on Jacobi updates have been proposed. Then, it has been explained why and how the above method may be customized to separate specific correlated signals such as images sharing a same template; the corresponding algorithm was named NOSWICA. Furthermore, the method was extended to perform on source signals being possibly bounded on one-side only, to yield the LABICA algorithm. This algorithm was designed to perform on “hard ICA problems” (involving noisy mixtures with few sample points, generated from ill-conditioned large mixing matrix and mixtures). Generally speaking, range-based and absolute bounds-based techniques are expected to suffer from a lack of robustness in the presence of additive noise on the mixtures; some experiments have confirmed that drawback. However, the LABICA algorithm was tested on a competition benchmark: outstanding results were obtained compared to other well-known algorithms as well as compared to other dedicated techniques presented in the framework of the related competition. The method is proved to be highly robust to the dimensionality of the source space, to the condition number of the mixing matrix and still performs well when few data samples are available. Improving the robustness to mixture noise remains a challenging issue; that could be dealt with by using more sophisticated range estimators.

4.6 APPENDIX

4.6.1 Proof of relation (4.52)

The key idea to find $p_{R_N(X)}(r)$ is to first compute the joint pdf of the extreme samples, i.e. $g_{X_{(1:N)}, X_{(N:N)}}(u, v)$. Noting that if the i -th order statistics $X_{(i:N)}$ is used instead of the i -th largest value of a sample set of size N in the definition

of $R_N(\mathbf{X})$, $\mathbf{X}_{(N:N)} = R_N(\mathbf{X}) + \mathbf{X}_{(1:N)}$, we find:

$$p_{R_N(\mathbf{X})}(r) = \int_u g_{\mathbf{X}_{(1:N)}, \mathbf{X}_{(N:N)}}(u, r+u) du , \quad (4.50)$$

where the range of the above integration is over the possible values of u when r is fixed (that is such that $g_{\mathbf{X}_{(1:N)}, \mathbf{X}_{(N:N)}}(u, r+u)$ exists).

Let us now turn to $g_{\mathbf{X}_{(1:N)}, \mathbf{X}_{(N:N)}}(u, v)$. To this end, observe that the probability that each of the N (assumed i.i.d.) realizations of \mathbf{X} is in (u_0, v_0) is

$$\prod_{i=1}^N \Pr[u_0 \leq X_i \leq v_0] = [\Pr(X_i = v_0) - \Pr(X_i = u_0)]^N = \left[\int_{u_0}^{v_0} p_X(x) dx \right]^N . \quad (4.51)$$

On the other hand, since $u_0 \leq v_0$, one has

$$\Pr[X_{(1:N)} \geq u_0 \text{ & } X_{(N:N)} \leq v_0] = \int_{u=u_0}^{v_0} \int_{v=u}^{v_0} g_{\mathbf{X}_{(1:N)}, \mathbf{X}_{(N:N)}}(u, v) du dv .$$

Therefore, we have

$$-\int_{u=v_0}^{u_0} \int_{v=u}^{v_0} g_{\mathbf{X}_{(1:N)}, \mathbf{X}_{(N:N)}}(u, v) dv du = \left[-\int_{v_0}^{u_0} p_X(x) dx \right]^N .$$

Remind that u_0 and v_0 are arbitrary values in $\Omega(\mathbf{X})$. Let us first keep v_0 fixed and differentiate w.r.t u_0 before doing the opposite, we get, using the standard differential calculus:

$$g_{\mathbf{X}_{(1:N)}, \mathbf{X}_{(N:N)}}(u, v) = N(N-1)p_X(u)p_X(v) \left[\int_u^v p_X(x) dx \right]^{N-2} .$$

Finally, since $r = v - u$, one gets, putting $\Omega(\mathbf{X}) = (a, b)$:

$$\begin{aligned} p_{R_N(\mathbf{X})}(r) &= \int_u g_{\mathbf{X}_{(1:N)}, \mathbf{X}_{(N:N)}}(u, r+u) du \\ &= N(N-1) \\ &\quad \times \int_a^{b-r} p_X(u)p_X(u+r) \left[\int_u^{u+r} p_X(x) dx \right]^{N-2} du . \end{aligned} \quad (4.52)$$

4.6.2 Expectation of the order statistics cdf differences

In this appendix, we show that $E[\Pr(X_{(N:N)}) - \Pr(X_{(1:N)})] = \frac{N-1}{N+1}$. It is obviously the case when $P_X = P_U$ since we have shown that in this case $E[R_N(U)] = \frac{N-1}{N+1}$. We show that this result still holds whatever the pdf P_X .

Let us first give an original proof of the following (known) result.

Proposition 6 *Let \mathbf{X} be a r.v. drawn from pdf $p_X(x)$ and cdf $\Pr_X(x)$. Then, for $i \in [1, N-1]$ ($N \geq 2$) we have $E[\Pr(X_{(i+m:N)}) - \Pr(X_{(i:N)})] = \frac{m}{N+1}$ and*

$E[P_X(x_{(N-p+1:N)}) - P_X(x_{(p:N)})] = \frac{N-2p+1}{N+1}$, which does not depend on the pdf of X .

Proof: Let $U_{(i:N)} = P_X(X_{(i:N)})$ be the random variable corresponding to the cdf $P_X(x)$ value of X evaluated at a point x corresponding to the realization of $X_{(i:N)}$; this r.v. is uniformly distributed on $[0, 1]$. By using basic properties of expectation, we have:

$$E[U_{(i+1:N)} - U_{(i:N)}] = \int_0^1 x p_{U_{(i+1:N)}}(x) dx - \int_0^1 x p_{U_{(i:N)}}(x) dx . \quad (4.53)$$

On the other hand, it is known from [Rose and Smith, 2002] that if X is drawn from the pdf $p_X(x)$, then the pdf of the i -th order statistics $X_{(i:N)}$ of X is:

$$p_{X_{(i:N)}}(x) = \frac{N!}{(i-1)!(N-i)!} [P_X(x)]^{i-1} [1 - P_X(x)]^{N-i} p_X(x) . \quad (4.54)$$

Since here $U_{(i:N)} = P_X(X_{(i:N)})$, then $P_{U_{(i:N)}}(u) = u$ and $p_{U_{(i:N)}}(u) = 1$. Eq. (4.53) can be rewritten as:

$$\begin{aligned} E[U_{(i+1:N)} - U_{(i:N)}] &= \frac{N!}{i!(N-i-1)!} \int_0^1 x^{i+1} (1-x)^{N-i-1} dx \\ &\quad - \frac{N!}{(i-1)!(N-i)!} \int_0^1 x^i (1-x)^{N-i} dx \end{aligned} \quad (4.55)$$

Algebraic manipulations lead to

$$E[U_{(i+1:N)} - U_{(i:N)}] = \frac{1}{N+1} \frac{N!}{i!(N-i-1)!} \int_0^1 x^i (1-x)^{N-i-1} dx . \quad (4.56)$$

Applying iteratively the following formula [Spiegel, 1974]

$$\int x^m (ax+b)^n dx = \frac{x^m (ax+b)^{n+1}}{(m+n+1)a} - \frac{mb}{(m+n+1)a} \int x^{m-1} (ax+b)^n dx \quad (4.57)$$

shows that the integral in (4.56) equals $\frac{i!(N-i-1)!}{N!}$.

On the other hand,

$$E[P_X(x_{(N-p+1:N)}) - P_X(x_{(p:N)})] = \sum_{i=p}^{N-p} E[P_X(x_{(i+1:N)}) - P_X(x_{(i:N)})] , \quad (4.58)$$

which is equal to $\frac{N-2p+1}{N+1}$.

□

Setting $p = 1$, in the above Proposition, we find again $E[P_X(x_{(N:N)}) - P_X(x_{(1:N)})] = \frac{N-1}{N+1}$.

4.6.3 Variance of the order statistics cdf differences

Proposition 7 Let X be a r.v. drawn from pdf $p_X(x)$ and cdf $P_X(x)$, and $X_{(i:N)}$ be the i -th order statistics of an N -sampling of U . Then, for $i \in [1, N - 1]$ ($N \geq 2$) we have $\text{Var}[P_X(X_{(i+m:N)}) - P_X(X_{(i:N)})] = \frac{m(N+1-m)}{(N+2)(N+1)^2}$ and $\text{Var}[P_X(X_{(N-p+1:N)}) - P_X(X_{(p:N)})] = \frac{2p(N-2p+1)}{(N+2)(N+1)^2}$, which does not depend on the pdf of X .

Proof: Similar development as above on the p -th order statistics of a Uniform variable on $[0, 1]$ leads to

$$\text{Var}[U_{(i:N)}] = \frac{i(N-i+1)}{(N+2)(N+1)^2} = \text{Var}[U_{(N-i+1:N)}] \quad (4.59)$$

On the other hand, by using basic variance properties, we have that

$$\text{Var}[X - Y] = \text{Var}[X] + \text{Var}[Y] - 2\text{Cov}[X, Y] . \quad (4.60)$$

Since we know from [Papadatos, 1999, Szekely and Mori, 1985] that for rectangular (and thus uniform) pdf and $1 \leq i < j \leq N$:

$$\text{Corr}[U_{(i:N)}, U_{(j:N)}] = \sqrt{\frac{i(N+1-j)}{j(N+1-i)}} \quad (4.61)$$

we obtain

$$\text{Cov}[U_{(i+m:N)}, U_{(i:N)}] = \frac{i(N+1-i-m)}{(N+2)(N+1)^2} . \quad (4.62)$$

We find

$$\begin{aligned} \text{Var}[U_{(i+m:N)} - U_{(i:N)}] &= \text{Var}[U_{(i+m:N)}] + \text{Var}[U_{(i:N)}] - 2\text{Cov}[U_{(i+m:N)}, U_{(i:N)}] \\ &= \frac{m(N+1-m)}{(N+2)(N+1)^2} . \end{aligned} \quad (4.63)$$

The proposition is proven by using equations (4.60) and (4.63) setting $i = p$ and $m = N - 2p + 1$.

□

Setting $i = 1$ and $m = N - p$ in Proposition 7 gives

$$\text{Var}[(P_X(X_{(N:N)}) - P_X(X_{(1:N)}))] = \frac{2(N-1)}{(N+2)(N+1)^2} . \quad (4.64)$$

The last result is valid for all random variables X .

CHAPTER 5

CONCLUSION

In this thesis, an information-theoretic point of view is adopted for considering the blind source separation problem. In particular, the notion of “information measures” was the starting point of a possible generalized class of contrast functions. This unifying view leads us to the Rényi entropies. This is the purpose of Chapter 1, where some mathematical definitions and concepts are given as well. In the literature, some (functionals) of these entropies were proved or conjectured, separately, to be contrast functions for the linear instantaneous ICA problem and blind deconvolution. We have focused on the simplest mixture model (instantaneous, linear); the underlying motivation for doing that is twofold: i) it is also the most widely used BSS model and ii) some major issues have not been investigated even in this simple situation, so far.

We summarize below the main results of this thesis and point out new challenges and open questions arising from this work.

Summary of results

Chapter 2 aims at dissecting in minute details some of Rényi’s entropy properties in order to check if deflation, simultaneous or partial contrast functions can be built from these functionals. A positive answer was already given regarding e.g. the Shannon entropy when used in a simultaneous approach because in that case, Shannon’s entropy reduces to mutual information (up to a constant term) [Comon, 1994]. Also, other criteria are used even if the underlying motivation for using them remained unclear in many cases. Here, additional results

show that in a deflation scheme and under a unit-variance constraint on the output, exact Shannon's entropy reaches a local minimum point if the output is proportional to any of the non-Gaussian source, and (logically) a local maximum when the outputs correspond to the possible unique Gaussian source; this is stated in Theorem 11. The partial case was recently addressed in [Pham, 2006b]. Complementary results show that the zero-order Rényi entropy (called Hartley's entropy) also yields a contrast function for ICA; the condition about the number of non-Gaussian sources is replaced by the fact that the sources are supposed to be bounded, but some normalization constraint on the output must be kept in order not to converge to the trivial null signal. A very simple modification of this criterion yields a more appealing functional and therefore, we directly turn to this so-called "range-based" contrast. The opposite of the log-range can be seen to be an extended form of Rényi's entropy where r is set to zero. As for Shannon, the extended 0-Rényi entropy reaches a local minimum when the output is proportional to a source, still under a normalization constraint. This is proved for the simultaneous, deflation and partial separation schemes in Section 2.3.3, Theorem 14 and Theorem 16 (combined with Corollary 6), respectively. In addition, the deflation (partial) contrast function is proved to have a local minimum point when a (subset of) source(s) is recovered (Theorem 15 and Theorem 16).

On the contrary, it is proved in Section 2.4 that for *any* other value of the r Rényi exponent (that is for any $r > 0$, $r \neq 1$), some counter-examples can be found: in the simple $K = 2$ case involving two sources sharing the same pdf, the r -Rényi's entropy does not necessarily have a local minimum when a source is extracted; the existence of such a local minimum is, however, a necessary condition to ensure that the opposite of this functional is a contrast function, under the normalization constraint. The exponential family suffices to emphasize this drawback. This observation allows us to partially answer an important question in the field of information theory about the possible superadditivity of Rényi's entropy power; this is formally stated in Corollary 11. The popular quadratic entropy, which has been proposed several times as an ICA contrast function based on experimental results (see [Hild et al., 2001, Erdogmus et al., 2002a, Hild et al., 2006b]) is not, unfortunately an exception to this result. In spite of the literature in the area (see [Bercher and Vignat, 2002, Erdogmus et al., 2002b, 2004]), it is explained here why using Rényi's entropy is not a good idea for blind deconvolution, too.

Another problem is then tackled; the problem of spurious local optima of entropy-based contrast functions. Indeed, we cannot conclude from the above results that there is an *equivalence* between the set of local maximizers of a contrast and the set of corresponding non-mixing matrices. Therefore, we have no guarantee that mixing maximum points do not exist: adaptive optimization algorithms could be stuck in such spurious solution. This problem is known to occur when Shannon's entropy is used; this has been pointed out via simulation results by various researchers [Cardoso, 2000, Learned-Miller and Fisher III, 2003, Vrins and Verleysen, 2005b, Boscolo et al., 2004]. Surprisingly however, neither convincing explanations nor theoretical proofs of that fact were proposed.

Therefore, this issue is addressed in Chapter 3. The analysis is restricted to Shannon, Hartley and extended Hartley entropies and all other entropies are ruled out, as they have been proved to be the only Rényi entropies from which, generally speaking, contrast functions can be built. Among this restricted class, only the extended Hartley entropy yields a contrast that is proved not to have a local maximum when the output is not proportional to one of the (assumed bounded) sources. This kind of contrast functions, that has a *local* maximum point *if and only if* the outputs are proportional to distinct sources, is qualified of “discriminant” contrast function. The range-based contrast is proved to be discriminant whatever the extraction scheme, even without prewhitening and over the whole space of square matrices of same size as the source space (even though not over Jacobi trajectories, as shown by the counter example provided in Section 3.4.5, found via a preliminary theoretical study). To our knowledge, this is the first result of this kind. The discriminacy of the deflation contrast is established in Section 3.4.2 using a small variation approach, a geodesic convexity viewpoint or yet a second order analysis of the stationary point of the criterion. The corresponding results for the simultaneous and partial contrasts are given in Corollary 16 and Corollary 18, respectively. In these approaches, careful precautions are taken because the contrast function is not differentiable everywhere.

Conversely, Shannon’s and (regular) Hartley’s entropies do not benefit from this interesting discriminacy property. Counter-examples are given for Shannon and regular Hartley entropies involving multimodal source densities. Why multimodal? This question is addressed too, both from intuitive (Section 3.2.1) and formal (Section 3.2.2 and Section 3.2.3) points of view. These counter-examples are chosen according to theoretical and intuitive results. Hence, the range-based contrast is the only one in the generalized entropies family to be discriminant. However, another discriminant contrast function exists (in a deflation scheme [Delfosse and Loubaton, 1995] or, with $K = 2$, in simultaneous separation framework [Murillo-Fuentes and Gonzalez-Serrano, 2004]), based on the kurtosis. But, just like the range, it is proved to focus on the tails of the densities (this is obvious for the range, and intuitively explained in Section 3.2.4 for the cumulant-based contrast), contrarily to the non-zero Rényi’s entropies that are sensitive to the whole structure of the output densities. Note however that in these existing methods, a prewhitening is required and the search space is restricted to the space of orthogonal demixing matrices, contrarily to the range-based approach.

Due to the appealing properties of the range-based contrast, the range deserved to be analyzed under other viewpoints than the purely theoretical aspects. This is the goal of Chapter 4. In particular, various geometric interpretations are provided and the practical estimation of the range is studied in the specific framework of ICA. This leads to a simple but efficient deflation source separation algorithm (ICAforNDC and SWICA). A simple range estimator is proposed, which matches the ICA requirements in the sense that it is easy to evaluate and rather robust to a variation of the pdf shape; this is a critical point in BSS.

This estimator is analyzed under various viewpoints, but the underlying source separation application was always kept in mind. The good separation performance of this algorithm illustrates the efficiency and reliability of the method. Finally, another extension is proposed to deal with the non-orthogonal case, that is to work even without prewhitening (NOSWICA). It has been explained however that prewhitening might be useful in order to simplify the recovering of the sources (better conditioning of the mixtures). But this does not imply that the search space must be restricted to the set of orthogonal matrices. Indeed, even if theoretically speaking, the set of satisfactory demixing matrices belongs to the set of orthogonal matrices if a prewhitening step is performed (and actually, one can even freely assume that the demixing matrix is a *special* orthogonal matrix), a rigid orthogonalization constraint is not always desirable; it may lead to a cumulation of errors (which is a well-known weakness of orthogonal deflation approaches) and consequently to a sub-optimal solution if the sample-based prewhitening step is not perfectly achieved. Moreover, the relaxation of the rigid orthogonality constraint may have other advantages; in particular, a non-orthogonal version of the above algorithm is proved to succeed well on the separation of correlated images if some conditions are met regarding the type of source correlation. The advantages and limitations of the method are emphasized, too. Finally, an extension of the minimum-range based technique to sources that are bounded on one side only is suggested and tested on an IEEE competition benchmark. Our method outperforms the most popular ICA algorithms as well as those (tailored for this benchmark) that were submitted to the competition.

Forthcoming challenges and open questions

Even if this thesis gives (hopefully) elements of answer to some (hopefully again) interesting issues that remained unsolved, it also raises new questions. Those that seem the most challenging ones to us are listed below:

- The connection between class II strict additivity and the contrast function and discriminacy has been established (see Theorem 24). The r -th root of the r -th cumulant is class II r -subadditive (while the cumulants are strictly additive but not class II) [Pham, 2001b]. But since all the r -th order cumulants ($r \geq 3$) yield contrast functions [Comon, 1994], and in particular, the kurtosis yields a discriminant contrast function [Delfosse and Loubaton, 1995] (after a positive mapping like absolute value or an even power), just like the range does, are there fundamental connections between the strict additivity of the basis functional of the contrast function, a form of r -sub/sup additivity and the discriminacy properties? Is the discriminacy property of the r -th cumulant-based criteria preserved for other values of r ?
- In Chapter 3, sufficient conditions (basically: strong multimodality) ensuring that spurious Shannon entropy local minima exist have been given.

They might not be necessary, though. Only the emerged part of the iceberg might have been pointed out. Specifically, the multimodal nature of the source densities may be unnecessary to induce spurious local optima even if, once again, all the theoretical and experimental results showing the existence of such spurious optima involved several multimodal source densities, so far. Hence, a analysis similar to the one proposed here extended to unimodal densities would be informative. In other words, is it possible to fill the gap between [Boscolo and Roychowdhury, 2003] (which proves that when mixtures of two “sufficiently Gaussian” and symmetric sources are considered, the local minimum of output mutual information with respect to the mixing angle is unique) and the results of Chapter 3? If yes, how?

- This work addresses the characterization of the local optima of theoretical functionals. In practice however, one works with approximated versions of these quantities. Therefore, it would be interesting to analyze directly these empirical criteria. Note that this has been done by Hyvärinen to analyze the non-mixing maxima of some approximations of the negentropy [Hyvärinen, 1997, Comon, 1994]. In spite of our theoretical results, it is not surprising that illustrating the failure of some related algorithms (caused by spurious optima) is not an easy task; indeed, the plugged negentropy approximations are so strong that they may yield paradoxically to discriminant contrast functions ! Indeed, the squared kurtosis is one of these negentropy approximations [Hyvärinen et al., 2001] in the case of symmetric sources, and is well-known to lead to discriminant contrasts.
- Generally speaking, Rényi’s entropies are not contrast functions if no additional conditions are met. For example, Shannon’s entropy requires that at most one source has a Gaussian density (as usual when the source independence assumption is exploited); Hartley’s entropy assumes that the sources are bounded, such that their support measures are finite. We have shown that in general, the opposite of Rényi’s entropy with other values of the Rényi exponent is not a contrast function, because simple examples of sources densities are found to meet a necessary condition implying this non-contrast behavior. A further (but more complicated) step would consist in specifying a subset of source densities for which Rényi’s entropy is a contrast function for a given value of r . In other words, what are the *necessary and sufficient* conditions on the source densities ensuring that the r -Rényi entropy is a contrast function? However, this would make sense only if one can guess if these conditions are met *based on the mixture samples only* so that choosing a right value for r is possible; is that feasible? If no, then this would mean that the use of Rényi entropy for BSS should be definitively dismissed.
- A more philosophical question that would deserve to be addressed concerns intuitive justifications explaining why Shannon and Hartley entropies are

suitable functionals from the viewpoint of BSS, that is from the viewpoint of “complexity measure” as defined in this thesis. Some specificities of these entropies compared to the other Rényi entropies have been given in the literature [Aczel et al., 1974, Aczel and Daroczy, 1975, Knuth, 2005, Rényi, 1976b]. This is not *the panacea* however, as the connections between these results (that deal with discrete random processes only) and the fact that they lead to contrast function remains vague. Performing some research in that field might yield to a better understanding of contrast functions and to new separation criteria.

- Our results seem to scrap the intuition. Indeed, on the one hand it is known that a very *rough estimator* of Shannon’s entropy (the kurtosis) has no mixing maximum point (at least in the deflation scheme and under pre-whitening). On the other hand, it is proved here that when using the *exact* entropy-based criteria, spurious maximum points may exist in the corresponding contrast functions (and that a spurious solution may be found when gradient-ascent techniques are used, involving the exact or, by extension, well-approximated source score functions). Therefore, it is tempting to advise to prefer rough estimates of the entropy gradient than a precise one. However, other recent results indicate that the efficiency of (a variant of) the FastICA algorithm attains the Cramér-Rao bound if, briefly, the true score functions are the non-linearities plugged in FastICA [Koldovsky et al., 2006, Tichavsky et al., 2006]. Therefore, if precise results are needed, our advise is to use a two steps procedure which consists in 1) running a first ICA algorithm, maximizing a discriminant contrast function (i.e. possibly a rough entropy estimator) to be close to a good (non-mixing) solution, and afterwards 2) running a second ICA algorithm involving efficient estimates of true score functions (see e.g. [Pham, 2003] for score function estimation), in which the result of 1) is used as a good initial point. Is the computational time drawback of this procedure compensated by significantly better results ? Is it possible to merge these two steps to have a discriminant and simultaneously efficient algorithm ?
- How does our results generalize to other mixing schemes such as convolutive filtering? Some preliminary experimental results (not provided here for conciseness) show that the range criterion can be used to perform blind deconvolution (more precisely, for some mixing filters, we have been able to blindly invert them based on the minimum range approach). But are there other assumptions to ensure that it is a contrast function for blind deconvolution? What about the local optima problem and the discriminacy property in this case? We expect that local optima may exist, as Torkkola showed that when delays exist between the sources, numerous local attractors appear [Torkkola, 1996].
- Chapter 4 addresses the practical aspects of range-based source separation. In spite of its nice performance in terms of API, the convergence rate

should be improved, as well as the robustness to noise. Therefore, we think that some research should be done in this field. Alper Erdogan has recently proposed an alternative based on sub-differentials [Erdogan, 2006], but other solutions might be proposed. For instance, the g-convexity could be exploited (leading to possible *g-convex optimization* techniques) and other range estimators could be investigated more deeply. Similarly, other algorithms than Jacobi-like methods should be developed, as such optimization techniques may be stuck into spurious solutions due to the limited number of feasible trajectories.

REFERENCES

- P.A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. in preparation. xvi, 25, 27, 143
- S. Achard. *Mesure de Dépendance pour la Séparation Aveugle de Sources. Application aux Mélanges Post-Nonlinéaires*. Phd thesis, Université J. Fourier, Grenoble (France), 2003. 17, 165
- J. Aczel. On mean values. *Bulletin of the American Mathematical Society*, 54: 392–400, 1948. 40
- J. Aczel and Z. Daroczy. *On Measures of Information and Their Characterizations*. Academic Press, 1975. 42, 260
- J. Aczel, B. Forte, and C.T. Ng. Why the Shannon and Hartley entropies are “natural”? *Advances in Applied Probability*, 6:131–146, 1974. 77, 260
- S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society Series B (Methodological)*, 18(1):131–142, 1966. 14, 15
- L. Almeida and M. Faria. Separating a real-life non-linear mixture of images. In C.G. Puntonet and A. Prieto, editors, *Independent Component Analysis and Blind Signal Separation*, volume LNCS 3195 of *Lecture Notes in Computer Science*, pages 734–741, Granada (Spain), September 2004. Springer. Fifth ICA international conference. 227
- S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998. 25
- A. Antoniadis, A. Guérin-Dugué, C. Jutten, and D.-T. Pham, editors. *De la séparation de sources à l’analyse en composantes indépendantes: méthodes algorithmes et applications*, École de printemps, Villard-de-Lans (France), May 2001. 3
- C. Arndt. *Information Measures. Information and its Description in Science and Engineering*. Springer, Berlin Heidelberg, 2004. 37

- S. Artstein, K.M. Ball, F. Barthe, and A. Naor. Solution of Shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4):975–982, 2004. 56, 57
- M. Babaie-Zadeh. *On Blind Source Separation in Convulsive and Nonlinear Mixtures*. Phd thesis, Université J. Fourier, Grenoble (France), 2002. 166
- M. Babaie-Zadeh and C. Jutten. A general approach for mutual information minimization and its application to blind source separation. *Signal Processing*, 85(5):975–995, 2005. 167
- F.R. Bach and M.I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002. 17
- J. Balatoni and A. Rényi. On the notion of entropy. *Selected Papers of Alfred Rényi*, 1:558–586, 1976. 37
- A.R. Barron. Monotonic central limit theorem for densities. Technical Report 50, Department of Statistics, Stanford University, California (USA), 1984. 57
- A.R. Barron. Entropy and the central limit theorem. *Annals of Probability*, 14:336–342, 1986. 57
- F. Barthe. Optimal Young's inequality and its converse: a simple proof. *Geom. Funct. Anal.*, 8(2):234–242, 1998. 79
- M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, 1989. 14
- W. Beckner. Inequalities in Fourier analysis. *Bulletin of the American Mathematical Society*, 102:159–182, 1975. 80
- A. J. Bell and T. J. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995. 53
- M. Ben-Bassat and J. Raviv. Rényi's entropy and the probability of error. *IEEE Transactions on Information Theory*, 24(3):324–331, 1978. 41
- J.-F. Bercher and C. Vignat. A Rényi entropy convolution inequality with application. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2002. 79, 80, 81, 82, 256
- D. S. Bernstein. *Matrix Mathematics*. Princeton University Press, 1954. 152
- N. Blachman. The convolution inequality for entropy powers. *IEEE Transactions on Information Theory*, 11:267–271, 1965. 38
- Y. Blanco and S. Zazo. An overview of BSS techniques based on order statistics: Formulation and implementation issues. In C.G. Puntonet and A. Prieto,

- editors, *Independent Component Analysis and Blind Signal Separation*, volume LNCS 3195 of *Lecture Notes in Computer Science*, pages 73–80, Granada (Spain), September 2004. Springer. Fifth ICA international conference. 250
- R. Boscolo, Hong Pan, and V.W. Roychowdhury. Independent component analysis based on nonparametric density estimation. *IEEE Transactions on Neural Networks*, 15(1):55–65, 2004. 102, 256
- R. Boscolo and V. Roychowdhury. On the uniqueness of the minimum of the information-theoretic cost function for the separation of mixtures of nearly gaussian signals. In *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 137–141, Nara (Japan), 2003. 167, 259
- H.J. Brascamp and E.H. Lieb. Best constants in Youngs inequality, its converse, and its generalization to more than three functions. *Bulletin of the American Mathematical Society*, 20:151–173, 1976. 80
- J.-B. Brissaud. The meanings of entropy. *Entropy*, 7(1):68–96, 2005. 37
- M. Brookes. The matrix reference manual. Technical report, Imperial College, London (UK), 2005. available online at <http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html>. 115
- J.-C. Cardoso. *Unsupervised Adaptive Filtering*, volume 1, chapter Entropic Contrast for Source Separation:Geometry & Stability, pages 265–319. Wiley and Sons, New York, 2000. 15, 102, 256
- J.-F. Cardoso. Source separation using higher order moments. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2109–2112, Glasgow (England), 1989. 224
- J.-F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4(4):112–114, April 1997. 53
- J.-F. Cardoso. Blind signal separation: Statistical principles. In R.-W. Liu and L. Tong, editors, *Proceedings of the IEEE, Special issue on blind identification and estimation*, pages 2009–2025. IEEE, 1998. 17, 20
- J.-F. Cardoso. Dependence, correlation and gaussianity in independent component analysis. *Journal of Machine Learning Research*, 4:1177–1203, 2003. 15
- J.-F. Cardoso and P. Comon. Independent component analysis, a survey of some algebraic methods. In *Proceedings of the IEEE International Symposium Circuits and Systems (ISCAS)*, volume 2, pages 93–96, Atlanta (USA), 1996. 28
- J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEEE proceedings F*, 140(6):362–370, 1993. 17, 28, 224

- C. Chefd'Hotel, D. Tschumperlé, R. Deriche, and O. Faugeras. Regularizing flows for constrained matrix-valued images. *Journal of Mathematical Imaging and Vision*, 20:147–162, January 2004. 25
- J.T. Chu. Some uses of quasi-ranges. *Annals of Mathematical Statistics*, 28(28):173–180, 1957. 218
- A. Cichocki and P. Georgiev. Blind source separation algorithms with matrix constraints. *IEICE Transactions on Fundamentals*, E-86-A(1):1–9, January 2003. 227, 228
- A. Cichocki and S.-I. Amari. *Adaptive blind signal and image processing*. John Wiley ans Sons, England, 2002. 3
- P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. 7, 17, 20, 46, 52, 136, 164, 255, 258, 259
- P. Comon. From source separation to blind equalization, contrast-based approaches. In *Proceedings of the International Conference on Image and Signal Processing (ICISP)*, Agadir, Morocco, May 2001. 138
- P. Cooke. Statistical inference for bounds of random variables. *Biometrika*, 66:367–374, 1979. 199
- P. Cooke. Optimal linear estimation of bounds of random variable. *Biometrika*, 67(1):257–258, 1980. 199
- M. Costa and Th. Cover. On the similarity of the entropy power inequality and the Brunn-Minkowski inequality. *IEEE Transactions on Information Theory*, 6(30):837–839, 1984. 60, 77
- T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley and Sons, 1991. 15, 30, 32, 36, 40, 60, 70, 85
- H. Cramér. *Mathematical Methods in Statistics*. Princeton University Press, New Jersey, 1946. 54
- S. Cruces, A. Cichocki, and S. Amari. The minimum entropy and cumulants based contrast functions for blind source extraction. In J. Mira and A. Prieto, editors, *Connectionist Models of Neurons, Learning Processes and Artificial Intelligence*, volume LNCS 2084 of *Lecture Notes in Computer Science*, pages 786–793, Granada (Spain), 2001. Springer-Verlag. Sixth International Work-conference on Artificial Neural Networks (IWANN). 59
- S. Cruces, A. Cichocki, and S. Amari. From blind signal extraction to blind instantaneous signal separation: criteria, algorithms and stability. *IEEE Transactions on Neural Networks*, 15(4):859–873, July 2004. 47, 59
- S. Cruces and I. Duran. The minimum support criterion for blind source extraction: a limiting case of the strengthened Young's inequality. In C.G.

- Puntonet and A. Prieto, editors, *Independent Component Analysis and Blind Signal Separation*, volume LNCS 3195 of *Lecture Notes in Computer Science*, pages 57–64, Granada (Spain), September 2004. Springer. Fifth ICA international conference. 46, 62, 65
- I. Csiszar. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967. 14
- G. Darmois. Analyse générale des liaisons stochastiques. *Revue de l’Institution Internationale de Statistique*, 21:2–18, 1953. 12
- H.A. David. *Order Statistics*. Wiley, New York, 1970. 205
- A. Delaigle and I. Gijbels. Boundary estimation and estimation of discontinuity points in deconvolution problems. Discussion paper 0325, Université catholique de Louvain, Institute of Statistics, 2003. 201
- N. Delfosse and P. Loubaton. Adaptive blind separation of sources: a deflation approach. *Signal Processing*, 45:59–83, 1995. 135, 164, 257, 258
- A. Dembo. Information inequalities and uncertainty principles. Technical report, Department of Statistics, Stanford University, California (USA), 1990. 60
- A. Dembo, T. M. Cover, and J. A. Thomas. Information theoretic inequalities. *IEEE Transactions on Information Theory*, 37(6):1501–1518, 1991. 80
- L. Devroye and L. Györfi. *Nonparametric Density Estimation*. Wiley, New York, 1985. 202
- L. Devroye and G.L. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal of Applied Mathematics*, 38:480–488, 1980. 201
- A.T. Erdogan. A simple geometric blind source separation method for bounded magnitude sources. *IEEE Transactions on Signal Processing*, 54:438–447, February 2006. 223, 239, 261
- D. Erdogmus, K.E. Hild, and J.C. Principe. Blind source separation using Rényi’s α -marginal entropies. *Neurocomputing*, 49(49):25–38, 2002a. 46, 66, 79, 256
- D. Erdogmus, K.E. Hild, J.C. Principe, M. Lazaro, and I. Santamaria. Adaptive blind deconvolution of linear channels using Rényi’s entropy with Parzen window estimation. *IEEE Transactions on Signal Processing*, 52(6):1489–1498, June 2004. 79, 80, 81, 256
- D. Erdogmus, J.C. Principe, and L. Vielva. Blind deconvolution with minimum Rényi’s entropy. In *Proceedings of the European Signal Processig Conference (EUSIPCO)*, volume 2, pages 71–74, Toulouse (France), 2002b. 79, 80, 81, 256

- J. Even. *Contributions à la Séparation de Sources à l'Aide de Statistiques d'Ordre*. Phd thesis, Université J. Fourier, Grenoble (France), 2003. 250
- W. Feller. *An Introduction to Probability Theory and its Applications*, volume 2. John Wiley and Sons, Inc., New York, 1966. 12, 54, 61, 105, 199, 211
- P. Flandrin, R. Baraniuk, and O. Michel. Time-frequency complexity and information. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, Adelaide, Australia, 1994. 66
- J.H. Friedman. Exploratory projection pursuit. *Journal of the American Statistics Association*, 82(397):249–266, 1987. 136
- R.J. Gardner. The Brunn-Minkowski inequality. *Bulletin of the American Mathematical Society*, 3(39):355–405, 2002. 60, 79, 80
- M. Gell-Mann and J. Hartle. *Physical Origins of Time Asymmetry*, chapter Time Symmetry and Asymmetry in Quantum Mechanics and Quantum Cosmology. Cambridge University press, Cambridge (England), 1996. xxiii
- R. Gray and L. Davisson. *An Introduction to Statistical Signal Processing*. Cambridge University Press, 2004. 32, 53, 69, 115
- R. M. Gray. *Entropy and Information Theory*. Springer-Verlag, New York, 1991. 32
- R.M. Gray, D.L. Neuhoff, and P.C. Shields. A generalization of Ornstein's \bar{d} distance with applications to information theory. *Annals of Probability*, 3(2):315–328, 1975. 15
- F. A. Graybill. *Matrices with Applications in Statistics*. Duxbury, 1983. 152
- B. Greene. *The Fabric of the Cosmos*. Alfred A. Knopf, New York, 2005. xxiii
- P. Hall. On estimating the endpoint of a distribution. *The Annals of Statistics*, 10(2):556–568, 1982. 200
- P. Hall and L. Simar. Estimating a changepoint, boundary or frontier in the presence of observation error. *Journal of American Statistics Association*, 97(458):523–534, 2002. 201
- G.H. Hardy, J.A. Littlewood, and Ploya G. *Inequalities*. Cambridge, 1934. 40
- D. A. Harville. *Matrix Algebra from a Statistician's Perspective*. Springer, 1997. 152
- S. Haykin, editor. *Unsupervised Adaptive Filtering : Blind Source Separation*, volume 1. John Wiley and Sons, New York, 2000. 66

- K. E. Hild, D. Erdogmus, K. Torkkola, and J.C. Principe. Feature extraction using information-theoretic learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1385–1392, 2006a. 66
- K.E. Hild, D. Erdogmus, and J.C. Principe. Blind source separation using Renyi's mutual information. *IEEE Signal Processing Letters*, 8(6):174–176, June 2001. 66, 79, 256
- K.E. Hild, D. Erdogmus, and J.C. Principe. An analysis of entropy estimators for blind source separation. *Signal Processing*, 86:174–176182–194, 2006b. 66, 79, 256
- I.I. Hirschman and D.V. Widder. *The Convolution Transform*. Princeton University Press, New Jersey, 1955. 61, 105
- P.G. Hoel. *Introduction to Mathematical Statistics*. John Wiley and Sons, New York, 1975. 205
- P.J. Huber. Projection pursuit. *Annals of Statistics*, 13(2):435–475, 1985. 50
- A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Processing Systems*, pages 273–279. The MIT Press, 1997. Tenth International Conference on Neural Processing Information Systems (NIPS). 46, 259
- A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, New York, 2001. 3, 4, 5, 53, 136, 233, 259
- A. Hyvärinen and E. Oja. A fast fixed-point algorithm independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997. 136, 224
- R. Jenssen, K. E. Hild, D. Erdogmus, J. C. Principe, and T. Eltoft. Clustering using Renyi's entropy. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 523–528, Portland (USA), 2003. 66
- D. Johnson and S. Sinanovic. Symmetrizing the Kullback-Leibler distance. Technical report, Department of Electrical and Computer Engineering, Rice University, Texas (USA), March 2001. 15
- C. Jutten and J. Hérault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991. 17
- K.H. Knuth. Lattice duality: The origin of probability and entropy. *Neurocomputing* 67, 67:245–274, 2005. 77, 260
- Z. Koldovsky, P. Tichavsky, and E. Oja. Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. *IEEE Transactions on Neural Networks*, 17(5):1265–1277, 2006. 260

- S. Kullback. *Information Theory and Statistics*. Wiley and sons, 1959. 15
- S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. 15, 37
- E. G. Learned-Miller and J. W. Fisher III. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003. 102, 135, 256
- J.A. Lee. *Analysis of High-dimensional Numerical Data: From Principal Component Analysis to Non-linear Dimensionality Reduction and Blind Source Separation*. Phd thesis, Université catholique de Louvain, Louvain-la-Neuve (Belgium), 2003. 25, 26
- J.A. Lee, F. Vrins, and M. Verleysen. A simple ICA algorithm for non-differentiable contrasts. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages cr1412.1–4, Antalya (Turkey), 2005. 222
- J.A. Lee, F. Vrins, and M. Verleysen. A least absolute bound approach to ICA - application to the MLSP 2006 competition. In *Proceedings of the IEEE workshop on Machine Learning for Signal Processing (MLSP)*, pages 41–46, Maybooth (Ireland), 2006a. 238, 242
- J.A. Lee, F. Vrins, and M. Verleysen. Non-orthogonal support-width ICA. In M. Verleysen, editor, *Advances in Computational Intelligence and Learning*, pages 351–358, Bruges (Belgium), April 2006b. Fourteenth European Symposium on Artificial Neural Networks (ESANN). 234
- J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. 125, 172
- W.-Y. Loh. Estimating an endpoint of a distribution with resampling techniques. *The Annals of Statistics*, 12(4):1543–1550, december 1984. 203
- D.G. Luenberger. *Introduction to Linear and Non-linear Programming*. Addison-Wesley Publishing Company, Massachusetts, 1973. 145
- E. Lutwak, D. Yang, and G. Zhang. Cramér-Rao and moment-entropy inequalities for Rényi entropy and generalized Fisher information. *IEEE Transactions on Information Theory*, 51(2):473–478, 2005. 41
- J. Raviv M. Hellman. Probability of error, equivocation, and the Chernoff bound. *IEEE Transactions on Information Theory*, 16(4):368–372, 1970. 125, 172
- D. J.C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003. 30, 31
- M. Madiman and A.R. Barron. Generalized entropy power inequalities and monotonicity properties of information. *IEEE Transactions on Information Theory*, 2006. submitted. 56, 57

- A. Meister. Support estimation via moment estimation in presence of noise. *Statistics*, 40(3):259–275, June 2006. xvi, 200
- E. Mourier. études du choix entre deux lois de probabilités. *Comptes Rendus de l'Académie des Sciences de Paris*, 223:712–714, 1946. 37
- J. Murillo-Fuentes and F.J. Gonzalez-Serrano. A sinusoidal contrast function for the blind separation of statistically independent sources. *IEEE Transactions on Signal Processing*, 52(12):3459–3463, December 2004. 135, 164, 257
- P. Pajunen. Blind source separation of natural signals based on approximate complexity minimization. In *Independent Component Analysis and Blind Signal Separation*, pages 267–270, Aussois (France), 1999. Proceedings of the first ICA international conference. 233
- N. Papadatos. Upper bound for the covariance of extreme order statistics from a sample of size three. *Indian Journal of Statistics Series A, Part 2*, 61(61):229–240, 1999. 253
- E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962. 66, 100
- K.B. Petersen and M.S. Pedersen. The matrix cookbook. Technical report, Intelligent Signal Processing Group , Technical University of Denmark, Lyngby (Denmark), 2005. 152, 180, 181
- D.-T. Pham. Blind separation of instantaneous mixture of sources via an independent component analysis. *IEEE Transactions on Signal Processing*, 44(11):2768–2779, 1996. 114
- D.-T. Pham. Blind separation of instantenaous mixtrures of sources based on order statistics. *IEEE Transactions on Signal Processing*, 48(2):363–375, 2000. 46, 61, 91, 250
- D.-T. Pham. Contrast functions for blind separation and deconvolution of sources. In T.W. Lee, T.P. Jung, S. Makeig, and .T.J. Sejnowski, editors, *Independent Component Analysis and Blind Signal Separation*, pages 37–42, San Diego (USA), 2001a. Third ICA international conference. 50
- D.-T. Pham. Contrast functions for ICA and source separation. Technical report of the BLISS project, Laboratoire de modélisation et calcul (IMAG-CNRS), Grenoble (France), May 2001b. 258
- D.-T. Pham. Mutual information approach to blind separation of stationary sources. *IEEE Transactions on Information Theory*, 48(7):1935–1946, 2002. 57
- D.-T. Pham. Fast algorithm for estimating mutual information, entropies and score functions. In *ICA 2003*, pages 17–22, April 2003. 260

- D.-T. Pham. Entropy of a variable slightly contaminated with another. *IEEE Signal Processing Letters*, 12(7):536–539, 2005. 57, 67, 68, 69
- D.-T. Pham. Blind partial separation of instantaneous mixtures of sources. In J. Rosca, D. Erdogmus, J.C. Principe, and S. Haykin, editors, *Independent Component Analysis and Blind Signal Separation*, volume LNCS 3889 of *Lecture Notes in Computer Science*, pages 868–875, Charleston SC, USA, March 2006a. Springer. Sixth ICA international conference. 46, 50
- D.-T. Pham. Contrasts for blind partial extraction of instantaneous mixtures of sources. *IEEE Signal Processing Letters*, 2006b. submitted. 19, 58, 256
- D.-T. Pham and C. Jutten, editors. *Analyse en composantes indépendantes et séparation de signaux*, Journés AS Séparation de sources et GdR ISIS, ENST (Paris), June 2003. 3
- D.-T. Pham and F. Vrins. Local minima of information-theoretic criteria in blind source separation. *IEEE Signal Processing Letters*, 12(11):788–791, 2005. 112
- D.-T. Pham and F. Vrins. Discriminacy of the minimum range approach to the simultaneous blind separation of bounded sources. In M. Verleysen, editor, *Advances in Computational Intelligence and Learning*, pages 377–382, Bruges (Belgium), April 2006. Fourteenth European symposium on artificial neural networks (ESANN). 150, 155
- D.-T. Pham, F. Vrins, and M. Verleysen. Spurious entropy minima for multi-modal source separation. In *Proceedings of the International Symposium on Signal Processing and Applications (ISSPA)*, pages 37–40, Sidney (Australia), 2005. 112, 121
- M. Plumley. Lie group methods for optimization with orthogonality constraints. In C.G. Puntonet and A. Prieto, editors, *Independent Component Analysis and Blind Signal Separation*, volume LNCS 3195 of *Lecture Notes in Computer Science*, pages 1245–1252. Springer-Verlag, 2004. Fifth ICA international conference. 25, 29
- H.V. Poor. Robust decision design using a distance criterion. *IEEE Transactions on Information Theory*, 26(5):575–587, 1980. 15
- A. Prieto, C.G. Puntonet, and B. Prieto. A neural learning algorithm for blind separation of sources based on geometric properties. *Signal Processing*, 64:315–331, 1998. 189
- J.C. Principe, D. Xu, and J.W. Fisher III. *Unsupervised Adaptive Filtering*, volume 1 of *Wiley Series on Adaptive Learning Systems for Signal Processing, Communications, and Control*, chapter Information-Theoretic Learning, pages 265–319. Wiley and Sons, Inc., New York, 2000. 79

- M.H. Quenouille. Approximate tests of correlation in time series. *Journal of the Royal Statistical Society (B)*, 11:68–84, 1949. 199
- T. Rapcsak. Geodesic convexity in non-linear optimization. *Journal of Optimization Theory and Applications*, 69:169–183, 1991. 143, 144, 145
- H. Reinchenbach. *The Direction of Time*. Dover publication, New York, 1984. xxiii
- A. Rényi. *Calcul des Probabilités*. Dunod, 1966. 30, 31, 54
- A. Rényi. On measures of entropy and information. *Selected papers of Alfred Rényi*, 2:565–580, 1976a. 40
- A. Rényi. Some fundamental questions of information theory. *Selected Papers of Alfred Rényi*, 2:526–552, 1976b. 77, 260
- C. Rose and M.D. Smith. *Mathematical Statistics with Mathematica*, chapter Order Statistics. Springer-Verlag, New York, 2002. 252
- P. Sahoo, C. Wilkins, and J. Yeager. Threshold selection using Rényi’s entropy. *Pattern Recognition*, 30:71–84, 1997. 66
- C. E. Shannon. A mathematical theory of communication. *Bell Systems Technology Journal*, 27:379–423 and 623–656, 1948. 32, 38
- B. W. Silverman. *Density Estimation*. Chapman, Hall/CRC (London), 1986. 202
- M.R. Spiegel. *Mathematical Handbook of Formulas and Tables*. McGraw-Hill, New York, 1974. 252
- A. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959. 38
- G.J. Szekely and T.F. Mori. An extremal property of rectangular distributions. *Statistics and Probability Letters*, 3:107–109, 1985. 253
- F. Theis. *Mathematics in Independent Component Analysis*. Phd thesis, Universität Regensburg, Regensburg (Germany), 2002. 12
- P. Tichavsky, Z. Koldovsky, and E. Oja. Performance analysis of the fastica algorithm and cramér-rao bounds for linear independent component analysis. *IEEE Transactions on Signal Processing*, 54(4):1189–1203, 2006. 260
- K. Torkkola. Blind separation of delayed sources based on information maximization. In *Proceedings of the IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 6:3509–3512, 1996. 260
- S. Verdu and D. Guo. A simple proof of the entropy-power inequality. *IEEE Transactions on Information Theory*, 52(5):2165–2166, 2006. 38, 56

- F. Vrins, C. Archambeau, and M. Verleysen. Entropy minima and distribution structural modifications in blind separation of multi-modal sources. In R. Fisher, R. Preuss, and U. von Toussaint, editors, *Proceedings of the International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt)*, AIP Conference proceedings (AIP 735), pages 589–596. American Institute of Physics, Melville, New York, 2004. 102, 107
- F. Vrins, C. Jutten, D. Erdogmus, and M. Verleysen. Zero-entropy minimization for blind extraction of bounded sources (BEBS). In J. Rosca, D. Erdogmus, J.C. Principe, and S. Haykin, editors, *Independent Component Analysis and Blind Signal Separation*, volume LNCS 3889 of *Lecture Notes in Computer Science*, pages 747–754, Charleston SC (USA), March 2006. Springer. Sixth ICA international conference. 60, 163
- F. Vrins, C. Jutten, and M. Verleysen. SWM : A class of convex contrasts for source separation. In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages V.161–V.164, Philadelphia (USA), March 2005a. 65, 139
- F. Vrins, J.A. Lee, and M. Verleysen. Can we always trust entropy minima in the ica context ? In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, volume cr1107, pages 1–4, Antalya (Turkey), 2005b. 112
- F. Vrins, J.A. Lee, and M. Verleysen. Filtering-free blind separation of correlated images. In A. Prieto J. Cabestany and F. Sandoval, editors, *Computational Intelligence and Bioinspired Systems*, volume LNCS 3512 of *Lecture Notes in Computer Science*, pages 1091–1099, Barcelona (Spain), 2005c. Springer. Eighth International Work-Conference on Artificial Neural Networks (IWANN). 228
- F. Vrins, J.A. Lee, and M. Verleysen. A minimum range approach to blind extraction of bounded sources. *IEEE Transactions on Neural Networks*, 2007a. In press. 62, 140
- F. Vrins and D.-T. Pham. Minimum range approach to blind partial simultaneous separation of bounded sources. *Neurocomputing*, 2007. Invited paper, in press. 64, 152
- F. Vrins, D.-T. Pham, and M. Verleysen. Mixing and non-mixing local minima of the entropy contrast for blind source separation. *IEEE Transactions on Information Theory*, 2007b. In press. 55, 121
- F. Vrins and M. Verleysen. Information theoretic vs cumulant-based contrasts for multimodal source separation. *IEEE Signal Processing Letters*, 12(3):190–193, 2005a. 138
- F. Vrins and M. Verleysen. On the entropy minimization of a linear mixture of variables for source separation. *Signal Processing*, 85(5):1029–1044, 2005b. 102, 107, 256

- F. Vrins and M. Verleysen. Minimum support ICA using order statistics. part i: Performance analysis. In J. Rosca, D. Erdogmus, J.C. Principe, and S. Haykin, editors, *Independent Component Analysis and Blind Signal Separation*, volume LNCS 3889 of *Lecture Notes in Computer Science*, pages 262–269, Charleston SC (USA), March 2006a. Springer. Sixth ICA international conference. 220
- F. Vrins and M. Verleysen. Minimum support ICA using order statistics. part ii: Performance analysis. In J. Rosca, D. Erdogmus, J.C. Principe, and S. Haykin, editors, *Independent Component Analysis and Blind Signal Separation*, volume LNCS 3889 of *Lecture Notes in Computer Science*, pages 270–277, Charleston SC (USA), March 2006b. Springer. Sixth ICA international conference. 220, 226
- H.H. Yang and S.-I. Amari. Adaptive on-line learning algorithms for blind separation-maximization entropy and minimum mutual information. *Neural Computation*, 9(7):1457–1482, 1997. 227
- R. Zamir and M. Feder. A generalization of the entropy power inequality with applications. *IEEE Transactions on Information Theory*, 39(5):1723–1728, 1993. 57

APPENDIX A

ANNOUNCEMENT OF THE IEEE MLSP 2006 DATA ANALYSIS COMPETITION

Large Scale, Ill-Conditioned Independent Component Analysis With Limited Number Of Samples

organizers: Andrzej Cichocki & Deniz Erdogmus.¹

We assume the standard linear model described in matrix form as: $\mathbf{X} = \mathbf{AS}$, where \mathbf{X} is the $K \times N$ matrix representing K observations for N consecutive time instants, \mathbf{A} is an $K \times K$ nonsingular mixing matrix and \mathbf{S} is an $K \times N$ matrix representing sources; K is the number of sources (equal to the number of sensors) and N is the number of available samples.

It is assumed that only the matrix \mathbf{X} is available, while the matrices \mathbf{A} and \mathbf{S} are unknown and should be estimated. The objective of this problem is to investigate the effect of increasing dimensionality n , decreasing number of samples T , increasing ill-conditioning of the mixing matrix \mathbf{A} and/or increasing the level of additive noise in the sensor level for performance and reliability of independent component analysis algorithms.

The original non-negative source signals are to be generated in MATLAB as follows:

```
for k = 1 : K :  
    if rand <= 0.5, S(k, :) = rand(1, N); % (sub-Gaussian)  
    else S(k, :) = -log(rand(1, N)).*max(0, sign(rand(K, N)) - 0.5); % (super-Gaussian)  
    end
```

¹Some symbol definitions have been modified in order to match the convention of the present thesis.

end

The mixing matrices are to be generated in MATLAB as follows: $\mathbf{A} = \text{rand}(K) - 0.5$;

The performance is measured using the average signal-to-interference measure (SIR) which is calculated in MATLAB using: $\mathbf{W} = \mathbf{B}\mathbf{A}$; $\text{SIR} = \text{mean}(10 * \log_{10}(\max(\mathbf{W}^2, [], 2). / (\sum(\mathbf{W}.*\mathbf{W}, 2) - \max(\mathbf{W}^2, [], 2))))$; where \mathbf{B} is the separation matrix such that $\mathbf{Y} = \mathbf{B}\mathbf{X}$, where \mathbf{Y} is the matrix of separated components.

The problem consists of four sub-problems. Performance of the competing algorithms MUST be provided for ALL four sub-problems. The algorithms need not be new propositions, however, in the case of a tie, novel approaches will be announced as the winner.

1. *Large scale problem* Determine the largest mixture dimension K for which the algorithm achieves an SIR greater than 15dB using $N = 5000$ samples. The experiment must be conducted using Monte Carlo runs where for each run the sources and the mixing matrix are generated randomly as described above. The $\text{SIR} > 15\text{dB}$ condition must be satisfied in 90% of the Monte Carlo runs for a particular value of K . Results should be presented in a figure that shows the following curves: $\text{SIR}_{10\%}$ vs K , $\text{SIR}_{50\%}$ vs K , and $\text{SIR}_{90\%}$ vs K . In general $\text{SIR}_P\%$ is the maximum real number such that $P\%$ of the Monte Carlo SIR values for a particular K are greater than this number. Note that $\text{SIR}_{90\%}$ vs K should be the last curve to cross over the desired 15dB threshold as n increases.²
2. *Small training set problem* Determine the smallest number of samples N for which the algorithm achieves an SIR greater than 15dB for $K = 50$. The experiment must be conducted using Monte Carlo runs where for each run the sources and the mixing matrix are generated randomly as described above. The $\text{SIR} > 15\text{dB}$ condition must be satisfied in 90% of the Monte Carlo runs for a particular value of N . Results should be presented in a figure that shows the following curves: $\text{SIR}_{10\%}$ vs N , $\text{SIR}_{50\%}$ vs N , and $\text{SIR}_{90\%}$ vs N . $\text{SIR}_P\%$ is described similar to sub-problem 1.
3. *Highly ill-conditioned problem*

Determine the highest dimension K for which the algorithm achieves an SIR greater than 15dB for $N = 5000$. The experiment must be conducted using Monte Carlo runs where for each run the sources are generated randomly as described above. Only in this sub-problem, the increasingly ill-conditioned mixing matrix is generated randomly as $\mathbf{A} = \mathbf{R}\mathbf{H}\mathbf{R}^T$ where $\mathbf{H} = \text{hilb}(K)$ is the Hilbert matrix as generated in MATLAB and \mathbf{R} is a random rotation matrix generated as shown below:

$$\text{ind} = \text{randperm}(K); \theta = 2 * \pi * \text{rand}; i = \text{ind}(1); j = \text{ind}(2);$$

²There was a mistake here as $\text{SIR}_{90\%}$ is the first curve to cross the threshold.

$$\mathbf{R} = \text{eye}(K); \quad \mathbf{R}(i,i) = \cos(\theta); \quad \mathbf{R}(j,j) = \mathbf{R}(i,i); \quad \mathbf{R}(i,j) = \sin(\theta); \quad \mathbf{R}(j,i) = -\mathbf{R}(i,j);$$

The SIR > 15dB condition must be satisfied in 90% of the Monte Carlo runs for a particular value of K . Results should be presented in a figure that shows the following curves: SIR_{10%} vs K , SIR_{50%} vs K , and SIR_{90%} vs K .

4. Noisy mixture problem

Suppose that $\mathbf{X} = \mathbf{AS} + \mathbf{N}$, where \mathbf{N} is an $K \times N$ matrix representing additive spatially white Gaussian distributed noise. Determine the lowest SNR for which the algorithm achieves an SIR greater than 15dB for $K = 50$ and $N = 5000$. SNR is defined as the ratio of the average mixture power $\text{mean}(\mathbf{E}[\mathbf{X}_i^2])$, where the mean is over mixture channels, to the noise power σ^2 , converted to dB using $10\log_{10}(\text{mean}(\mathbf{E}[\mathbf{X}_i^2])/\sigma^2)$. The experiment must be conducted using Monte Carlo runs where for each run the sources and the mixing matrix are generated randomly as described above. The SIR > 15dB condition must be satisfied in 90% of the Monte Carlo runs for a particular value of SNR. Results should be presented in a figure that shows the following curves: SIR_{10%} vs SNR, SIR_{50%} vs SNR, and SIR_{90%} vs SNR. Notice that this measure does not consider the noise corruption levels at the separated outputs, rather it is only concerned with the performance of the (inverse) model estimation.

Remarks:

1. Sources are non-negative so alternative methods to ICA such as NMF (Non-negative Matrix Factorization) or ICA with non-negativity constraints can be also implemented and tested. The source distributions are assumed to be unknown, therefore, preset fine-tuning that matches the source distributions for optimal performance is not allowed.
2. The proposed algorithms should solve any sub-problem in reasonable computation time, say, in a few minutes on a typical PC (so that Monte Carlo runs can be run by the organizing committee using submitted Matlab codes if necessary).
3. Report the results in a document (*.doc or *.pdf), where a brief description of the algorithm and appropriate references, as well as experimental results demonstrating performance on the sub-problems are included. In your submission, please also include a self-contained Matlab script of your code that is in ready-to-run condition to replicate the Monte Carlo results/figures to facilitate the replication of results if required.
4. Regarding the performance index, the user can additionally use the recently provided Matlab package “BSS-EVAL”, which is a MATLAB toolbox to

compute reliably performance measures in (blind) source separation within an evaluation framework where the original sources are available as ground truth. Download page: http://www.irisa.fr/metiss/bss_eval/.

5. Upon request, Matlab function to generate the random mixing matrix for subproblem 3 is as follows (it Generates an $K \times K$ random mixing matrix \mathbf{A} that has the same eigenvalues as an $K \times K$ Hilbert matrix):

```
functionA = generateA(K)
H = hilb(K); ind = randperm(K); theta = 2 * pi * rand; i = ind(1);
j = ind(2);
R = eye(K); R(i,i) = cos(theta); R(j,j) = R(i,i); R(i,j) =
sin(theta); R(j,i) = -R(i,j);
A = RHR^T.
```

APPENDIX B

AUTHOR'S PUBLICATION LIST

International Journal Papers

- JP1 On the Entropy Minimization of a Linear Mixture of Variables for Source Separation. **F. Vrins** & M. Verleysen, *Signal Processing* **85**(5), Elsevier, pp. 1029-1044, 2005.
- JP2 Information Theoretic vs Cumulant-Based Contrasts for Multimodal Source Separation. **F. Vrins** & M. Verleysen, *Signal Processing Letters* **12**(3), IEEE, pp. 190-193, 2005.
- JP3 Local Minima of Information-Theoretic Criteria in Blind Source Separation. D.-T. Pham & **F. Vrins**, *Signal Processing Letters* **12**(11), IEEE, pp. 788-791, 2005.
- JA1 Mixing and Non-Mixing Local Minima of the Entropy Contrast for Blind Source Separation. **F. Vrins**, D.-T. Pham & M. Verleysen, *Transactions on Information Theory*, IEEE (sub. December 2005, rev. October 2006 & November 2006, in press; expected publication issue: March 2007).
- JA2 A Minimum-Range Approach to Blind Extraction of Bounded Sources. **F. Vrins**, J.A. Lee & M. Verleysen, *Transactions on Neural Networks*, IEEE (sub. January 2006, rev. July 2006 & October 2006, in press; expected publication issue: March 2007).
- JA3 A Minimum-Range Approach to Blind Partial Simultaneous Separation of Bounded Sources. **F. Vrins** & D.-T. Pham, *Neurocomputing*, Elsevier (sub. July 2006, rev. September 2006, invited paper, in press; expected issue: Spring 2007).
- JS1 On the risk of using Rényi's entropy for blind source separation. D.-T. Pham & **F. Vrins**. In preparation, to be submitted.
- JTBS2 Extension of [ICP12]. J.A. Lee, **F. Vrins** & M. Verleysen . *Invited paper in Neurocomputing, in preparation.*

International Conference Papers published as collective book

- ICB1 Entropy Minima and Distribution Structural Modifications in Blind Separation of Multi-modal Sources. **F. Vrins**, C. Archambeau & M. Verleysen. In R. Fisher et al. (Eds), *Proc. of the Int'l Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering* (Max-Ent), Garching (Germany), July 25-30, 2004. AIP Conference proceedings **735**, American Institute of Physics, pp. 589-596.
- ICB2 Sensor Array and Electrode Selection for Non-Invasive Fetal Electrocardiogram Extraction by Independent Component Analysis. **F. Vrins**, C. Jutten & M. Verleysen. In C.G. Puntonet and A. Prieto (eds), *Proc. of the Int'l Conf. Independent Component Analysis and Blind Signal Separation* (ICA), Granada (Spain), September 22-24, 2004. Lecture Notes in Computer Science **3195**, Springer, pp. 1017-1024.
- ICB3 Filtering-Free Blind Separation of Correlated Images. **F. Vrins**, J.A. Lee and M. Verleysen. In J. Cabestany, A. Prieto and F. Sandoval (eds), Computational Intelligence and Bioinspired Systems, *Proc. of the Int'l Work-Conf. Artificial Neural Networks* (IWANN), Barcelona (Spain), June 8-10, 2005. Lecture Notes in Computer Science **3512**, Springer, pp. 1091-1099.
- ICB4 Minimum Support ICA Using Order Statistics. Part I: Quasi-Range Based Support Estimation. **F. Vrins** & M. Verleysen. In J. Rosca, D. Erdogmus, J.C. Principe and S. Haykin (eds), *Proc. of the Int'l Conf. Independent Component Analysis and Blind Signal Separation* (ICA), Charleston SC (USA), March 5-8, 2006. Lecture Notes in Computer Science **3889**, Springer, pp. 262 - 269.
- ICB5 Minimum Support ICA Using Order Statistics. Part II: Performance Analysis. **F. Vrins** & M. Verleysen. In J. Rosca, D. Erdogmus, J.C. Principe and S. Haykin (eds), *Proc. of the Int'l Conf. Independent Component Analysis and Blind Signal Separation* (ICA), Charleston SC (USA), March 5-8, 2006. Lecture Notes in Computer Science **3889**, Springer, pp. 270 - 277.
- ICB6 Zero-Entropy Minimization for Blind Extraction of Bounded Sources (BEBS). **F. Vrins**, C. Jutten, D. Erdogmus & M. Verleysen. In J. Rosca, D. Erdogmus, J. C. Prncipe and S. Haykin (eds), *Proc. of the Int'l Conf. Independent Component Analysis and Blind Signal Separation* (ICA), Charleston SC (USA), March 5-8, 2006. Lecture Notes in Computer Science **3889**, Springer, pp. 747 - 754.
- ICB7 Electrode Selection for Non-invasive Fetal Electrocardiogram Extraction using Mutual Information Criteria. R. Sameni, **F. Vrins**, F. Parmenier, C. Hrail, V. Vigneron, M. Verleysen, C. Jutten & M.B. Shamsollahi. To appear in *Proc. of the Int'l Workshop on Bayesian Inference*

and Maximum Entropy Methods in Science and Engineering (MaxEnt), Paris (France), July 25-30, 2006. AIP Conference proceedings, American Institute of Physics.

International Conference Papers published as proceedings

- ICP1 Improving Independent Component Analysis Performances by Variable Selection. **F. Vrins**, J. A. Lee, M. Verleysen, V. Vigneron & C. Jutten), *Proc. of the IEEE Signal Processing Workshop on Neural Networks for Signal Processing (NNSP)*, pp. 359-368, Toulouse (France), September 17-19, 2003.
- ICP2 On the Extraction of the Snore Acoustic Signal by Independent Component Analysis. **F. Vrins**, V. Bouillon, J Deswert, D. Bouvy, J. A. Lee, C. Eugne & M. Verleysen, *Proc. of the IASTED Int'l Conf. Biomedical Engineering (BioMed)*, pp. 326-331, Innsbruck (Austria), February 16-18, 2004.
- ICP3 Abdominal Electrodes Analysis by Statistical Processing for Fetal Electrocardiogram Extraction. **F. Vrins**, V. Vigneron, C. Jutten & M. Verleysen, *Proc. of the IASTED Int'l Conf. Biomedical Engineering (BioMed)*, pp. 244-249, Innsbruck (Austria), February 16-18, 2004.
- ICP4 Towards a Local Separation Performances Estimator using Common ICA Contrast Functions? **F. Vrins**, C. Archambeau & M. Verleysen, *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, pp. 211-216, Bruges (Belgium), April 28-30, 2004.
- ICP5 Flexible and Robust Bayesian Classification by Finite Mixture Models. C. Archambeau, **F. Vrins** & M. Verleysen, *Proc. Eur. Symp. Artificial Neural Networks (ESANN)*, pp. 75-80, Bruges (Belgium), April 28-30, 2004.
- ICP6 SWM : A Class of Convex Contrasts for Source separation. **F. Vrins**, C. Jutten & M. Verleysen, *Proc. of the IEEE Int'l Conf. Acoustics Speech and Signal Processing (ICASSP)*, pp. V.161-164, Philadelphia (USA), March 19-23, 2005.
- ICP7 Spurious Entropy Minima for Multimodal Source Separation. D.-T. Pham, **F. Vrins** & M. Verleysen, *Proc. of the Int'l Symp. Signal Processing and Applications (ISSPA)*, pp. 37-40, Sidney (Australia), August 28-31, 2005.
- ICP8 Can We Always Trust Entropy Minima in the ICA Context? **F. Vrins**, J.A. Lee & M. Verleysen, *Proc. of the Eur. Signal Processing Conf. (EUSIPCO)*, pp. cr1107.1-4, Antalya (Turkey), September 4-8, 2005.
- ICP9 A Simple ICA Algorithm for Non-Differentiable Contrasts. J.A. Lee, **F. Vrins** & M. Verleysen, *Proc. of the Eur. Signal Processing Conf. (EUSIPCO)*, pp. cr1412.1-4, Antalya (Turkey), September 4-8, 2005.

- ICP10 Discriminacy of the Minimum Range Approach to the Simultaneous Blind Separation of Bounded Sources. D.-T. Pham & **F. Vrins**, *Proc. of the Eur. Symp. Artificial Neural Networks* (ESANN), pp. 211-216, Bruges (Belgium), April 26-28, 2006.
- ICP11 Non-orthogonal Support Width ICA. J.A. Lee, **F. Vrins** & M. Verleysen, *Proc. of the Eur. Symp. on Artificial Neural Networks* (ESANN), pp. 211-216, Bruges (Belgium), April 26-28, 2006.
- ICP12 A Least Absolute Bound Approach to ICA - Application to the MLSP 2006 competition. J.A. Lee, **F. Vrins** & M. Verleysen, *Proc. of the Int'l Workshop on Machine Learning for Signal Processing* (MLSP), pp. 41-46, Maynooth (Ireland), September 5-8, 2006 (invited paper, winner of MLSP 2006 ICA competition).
- ICTBS13 Is the General Form of Rényi's Entropy a Contrast for Source Separation? **F. Vrins**, D.-T. Pham & M. Verleysen, *To be submitted*.

Other publications

1. Apprentissage par Projet en Électricité (partie 1): Conception et Réalisation d'un Système de Filoguidage Électromagnétique. **F. Vrins**, L. De Vroey, F. Labrique & C. Trullemans, *Coll. sur l'Enseignement des Technologies et des Sciences de l'Information et des Systèmes en Électronique, Électrotechnique et Automatique*, Toulouse (France), November 13-14.
2. Apprentissage par projet en électricité : exemples et mise en oeuvre. L. De Vroey, **F. Vrins**, D. Grenier, F. Labrique, C. Trullemans & C. Eugène, *J. sur l'Enseignement des Sciences et Technologies de l'Information et des Systèmes* (J3eA) **5**, 2006.
3. Valoriser l'investissement didactique des assistants de recherche au travers de colloques et de revues. **F. Vrins** & L. De Vroey, *Louvain Ingénieurs* **1**, AILv, March, pp. 4-7, 2004.
4. Information processing approaches for variable selection and source separation : application in multi-dimensional biomedical signal processing. **F. Vrins**, *DEA Thesis*, FSA, Université catholique de Louvain (Belgium).