The Minimum Risk Principle That Underlies the Criteria of Bounded Component Analysis

Sergio Cruces, Senior Member, IEEE, and Iván Durán-Díaz, Member, IEEE

Abstract—This work studies the problem of the blind extraction of a subset of bounded component signals from the observations of a linear mixture. In the first part of this article we analyze the geometric assumptions of the observations that characterize the problem, and their implications on the mixing matrix and latent sources. In the second part, we solve the problem by adopting the principle of minimizing the risk, which refers to the encoding complexity of the observations in the worst admissible situation. This principle provides an underlying justification of several bounded component analysis criteria, including the minimum normalized volume criterion of the estimated sources or the maximum negentropy-likelihood criterion with a uniform reference model for the estimated sources. This unifying framework can explain the differences between the criteria in accordance with their considered hypotheses for the model of the observations. This study is first presented for the case of the extraction of a complex and multidimensional source, and later is particularized for the case of the extraction of subsets of one-dimensional complex sources. The results also hold true in the case of real signals, where the obtained criteria for the extraction of a set of one-dimensional sources usually coincide with the existing bounded component analysis criteria.

Index Terms—Bounded component analysis, independent component analysis, source separation, negentropy-likelihood criterion, dependent sources.

I. INTRODUCTION

There are a large number of practical situations in which the relevant information cannot be acquired directly. Usually the desired information is concealed by the superposition of some nuisance components that are uninteresting for us. The additive decomposition of a multidimensional set of observations in terms of basic component signals of a simple structure is a problem with numerous applications. This has favored its research in several fields, like: statistics, signal processing, psychometry, geophysics, neuroscience, audio processing and communications.

The different solutions to this additive decomposition problem depend on the available knowledge, which includes the structure of the signal model and any implicit assumptions on the model parameters. When we can solve the problem without the need to resort to training data, the proposed methods can be considered blind or unsupervised. Our assumptions or hypotheses condition the best criteria and learning algorithms to choose for this decomposition. In absence of training samples, the blind criteria can drive the learning algorithms by optimizing one or several objective functions, which aim to summarize or condense those aspects of the data that are meaningful.

Unsupervised algorithms usually assume a simple algebraic structure for the desired components (typically a low-rank structure) together with a set of additional hypotheses, like the decorrelation and joint Gaussianity of the sources in factor analysis (FA), their mutual independence and non-Gaussianity in independent component analysis (ICA) [1], their non-negativity in nonnegative matrix factorization (NMF) [2], their sparsity in sparse component analysis (SCA) [3], or some geometric assumptions for the desired sources.

There is a large number of blind learning algorithms that have exploited the geometric properties of the component signals to solve the decomposition of the observations. In communications applications, blind deconvolution methods usually resort to properties that depend on the constellation of the transmitted symbols. One of the first examples was Godard’s family of criteria for the equalization of the intersymbol interference [4]. Other works proposed learning algorithms that explicitly exploit the square shape of the constellation [5], the constant modulus of the symbols [6] and the exact knowledge of the alphabet [7]. A good review of several geometric methods can be found in [8].

However, our study will focus on the specific class of methods that exploit a bounded assumption for the sources [9]-[20]. These methods have been proposed first in the field of independent component analysis and later in the specific framework of bounded component analysis. They have the attractive feature that in most cases the local optima of these criteria correspond with valid solutions of the extraction/separation problems. The pioneer technique was a minimum range criterion proposed by Pham in [9], which was based on the assumption of real and mutually independent sources of finite support and addressed the separation of all the sources. Another technique was the minimum support criterion, proposed in [10] for the extraction of a single real source from a mixture of bounded sources. The gap between these two techniques was closed with the extension of the method for the extraction of an arbitrary number of sources in [11] and [12], independently. Meanwhile, a different approach in [13] proposed the minimization of the maximum infinite norm of the real part of the outputs for sources with positive-negative peak symmetry.

In the framework of bounded component analysis (BCA) the independence assumption for the sources is relaxed [14]. A minimum normalized volume criterion for complex sources was proposed in [16] and the effectiveness of several BCA...
criteria in the extraction and separation of dependent sources was shown in [17]-[20].

There are several examples of the successful application of bounded component analysis in problems like: the separation of mixtures of independent and dependent sources in communications [13]-[20], the blind equalization [21] and identification [22] of communications channels, and the separation of mixtures of saturated pictures [23]. For observations with a high signal to noise ratio and when the desired sources have a clearly bounded structure, the BCA methods often obtain an excellent performance what motivates their preference over other criteria. In other situations, the independent component analysis [1] or alternative techniques for dependent sources [24] usually perform better.

In this work, we consider a minimum risk principle which is based on the minimization of the encoding complexity of the observations in the worse case admissible situation. Although this principle has been studied by several authors, among the many relevant contributions we can highlight those of Rissanen [25] and Topsøe [27]. The principle is related with the minimum information discrimination criterion proposed by Kullback [28] and also with the maximum entropy criterion of Jaynes [29]. Our implementation of the principle is inspired in the methods of projection pursuit density estimation [30] and their specific application in blind signal extraction [31]. The minimax mutual information criterion for independent component analysis [32] and the entropy bound minimization criterion [33], [34] can be also regarded as complementary interpretations of this underlying risk principle.

The main contribution of this article is to show how the bounded component analysis assumptions for the sources, together with this minimum risk principle for the estimation of the density of the observations, drive our best strategy towards the minimization of a bounded component analysis criterion. Depending on the specific hypotheses for the sources, the suitable evaluation of the minimum risk differs, which can explain the differences and similarities between the resulting BCA criteria.

This paper is organized as follows. The signal model of the observations is described in section II. In sections III and IV, we present the implications of the bounded component analysis assumptions on the components of the observations, the sources and the mixing system. We formulate in sections V and VI the results for the estimation of the density of the observations under different sets of BCA assumptions, while in sections VII and VIII, we discuss the unifying framework of the contribution. We present in section IX some illustrative simulations and summarize in section X our main conclusions.

For the reader’s convenience, table I presents a brief summary of the adopted notation.

II. DECOMPOSITION OF THE OBSERVATIONS IN TARGET AND NUISCANCE COMPONENTS

Consider that an \textit{m} dimensional and stationary data process is projected onto its signal subspace (of dimension \textit{n} \leq \textit{m}) and sampled to yield the set of the observation random vectors \{\textit{X}(\textit{t}), \textit{t} = 1, \ldots, \textit{T}\}. We assume that this is a set of independent and identically distributed (i.i.d.) random vectors, so for convenience from here on we will omit their temporal indices. We represent the observation random vector as \textit{X} = (X_1, \ldots, X_n)^T \in \mathbb{C}^n, where \textit{X}_i denotes the \textit{i}th random variable. For simplicity, we initially consider the decomposition of the observations in only two underlying components:

\begin{equation}
\textit{X} = \textit{X}_r + \textit{X}_\nu.
\end{equation}

The target or interesting component \textit{X}_r is assumed to be a bounded random vector, while, the complementary part of the additive decomposition \textit{X}_\nu represents a nuisance component which does not need to satisfy the previous assumption\footnote{Note that our definition of \textit{X}_\nu does not necessarily enforce this nuisance component to be unbounded. Indeed, when it is bounded, we have two bounded components in the mixture, so the roles of the target and nuisance components can be arbitrarily interchanged and, in general, we would need to resort to extra information in order to identify them.}.

Without loss of generality, we assume that the observations and their components are centered at the origin, so they are zero mean random vectors. Although this is not strictly necessary, it simplifies the treatment of possible degenerate (constant) components in the mixture which under this condition are always identified with the origin.

The additive decomposition in (1) is similar to the one considered in multidimensional independent component analysis [35], [36], except for the fact that we will no longer assume the mutual independence of the components. Let us denote by \textit{n} = \text{dim}(\textit{X}) the dimension of the space of the observations. A common but relevant assumption for the separability of the components, is that they belong to complementary subspaces, i.e. they only intersect at the origin. Thus, the space of the observations decomposes as a direct sum of the component subspaces,

\begin{equation}
\{\textit{X}\} = \{\textit{X}_r\} \oplus \{\textit{X}_\nu\}
\end{equation}

TABLE I

\begin{tabular}{|c|}
\hline
\textbf{TABLE I} & \textbf{SUMMARY OF THE NOTATION.} \\
\hline
\textbf{X} & random vector of the observations. \\
\textbf{X}_r & target component of the observations. \\
\textbf{X}_\nu & nuisance component of the observations. \\
\textbf{S}_X & convex support set of the random vector \textit{X}. \\
\textbf{p}_X & \text{density of } \textit{X} \text{ evaluated at } \textit{x}. \\
\textbf{\hat{p}}_X & \text{estimate for the density of } \textit{X}. \\
\textbf{P}_X & \text{set of admissible density functions of the observations.} \\
\textbf{\log} & \text{logarithm in base } 2 \text{ so as to measure information in bits.} \\
\textbf{X}_N & \text{complex Gaussian r.v. distributed as } \mathcal{CN}(\mu_X, \Sigma_X). \\
\textbf{X}^t & \text{uniform r.v. with support } \textit{S}_X. \\
\textbf{\textit{x}} & \text{sample realization of the random vector } \textit{X}. \\
\textbf{x}^{1:T} & \text{sample set of observations } \{\textit{x}(1), \ldots, \textit{x}(T)\}. \\
\textbf{n} & \text{dimensionality of the observations.} \\
\textbf{\textit{S}} & \text{dimensionality of the target source (}p \leq n\text{).} \\
\textbf{V}_{\text{det}}(\textit{x}) & \text{volume determined by the set } \textit{x}^{1:T}. \\
\textbf{\text{Conv}}_n(\textit{S}_X) & \text{volume of } \textit{S}_X \subseteq \mathbb{C}^n. \\
\langle \cdot \rangle & \text{subspace spanned by the columns of the argument.} \\
\times & \text{Cartesian product.} \\
\text{\text{\text{\oplus}}} & \text{direct sum of sets or subspaces.} \\
\text{|-|} & \text{for matrices denotes the modulo of the determinant.} \\
\end{tabular}
In particular, this enforces that the dimensions of the subspaces satisfy
\[
\dim(X) = \dim(X_r) + \dim(X_i)
\]
\[\frac{n}{n} = \frac{p}{p} + \frac{n-p}{n-p}\]  

Let us now introduce some basic notation and definitions from the fields of complex signal processing and convex geometry that we will find later useful for our purposes.

The complex random vector \( X = X_r + jX_i \in \mathbb{C}^n \) is determined by its real and imaginary parts \( X_r, X_i \in \mathbb{R}^n \). The probability density function (p.d.f) of the random vector \( X \) is defined as the joint density of its real and imaginary parts,
\[
p_X(x) = p(\mathbb{R}\{x\}, \Im\{x\}) \in \mathcal{P},
\]
where \( \mathcal{P} \) denotes the space of the valid probability density functions. The integral of a function \( f(x) \) with respect to the complex vector is interpreted as the double integral of the function with respect to the real and imaginary parts of the vector,
\[
\int f(x) dx \equiv \int \int f(x) d\mathbb{R}\{x\} d\Im\{x\}.
\]
This notation allows one to write most of the statistics of the complex random vector \( X \) is the following density function of its density
\[
\begin{align*}
\log h(p_X) & = \int p_X(x) \log \frac{1}{p_X(x)} dx, \\
\end{align*}
\]
where the logarithm is assumed to be in base 2 so as to yield this quantity in units of bits. This definition is complemented with its extension by continuity that enforces \( p_X(x) \log p_X(x) = 0 \) whenever \( p_X(x) = 0 \).

In the field of convex geometry it is usual to work with sets of real elements. However, with the definition for the integral in (5), it is quite simple to extend several tools from convex geometry for applying them to complex sets. The only caution is to remember that each complex dimension will account for a pair of real dimensions. In this way, for a complex set of \( T \) observations
\[
x_{1:T} = \{x(t), t = 1, \ldots, T\},
\]
with samples drawn independently from \( p_X(x) \), we define its closed convex hull as
\[
S_x = \cl \conv x_{1:T},
\]
which is the minimum closed convex set that encloses the \( T \) samples of the observations. Let us recall that the samples of the observations have intrinsic dimensionality \( n \), so the volume of \( S_x \) relative to the signal subspace (in this case \( \langle X \rangle = \mathbb{C}^n \)) is defined by
\[
\text{Vol}_{\mathbb{C}^n}(x) = V_{\mathbb{C}^n}(S_x) = \int_{\langle x \rangle} \mathbf{1}_{S_x} dx,
\]
where \( \mathbf{1}_{\langle x \rangle} \) denotes the indicator function.

Similarly, the convex support of the complex random vector \( X \), which we denote \( S_X \), is the closure of the convex hull of the set of points \( x \in \mathbb{C}^n \) for which the density of the random vector is nonzero, that is,
\[
S_X = \cl \conv \{x \in \mathbb{C}^n : p_X(x) > 0\}.
\]
For \( \lambda \in \mathbb{R}^+ \), a dilation by \( \lambda \) of the set \( S_X \) is given by
\[
\lambda S_X = \{\lambda x : x \in S_X\},
\]
and it is termed as the Minkowski product of \( S_X \) and \( \{\lambda\} \). For two sets \( S_{X_1} \) and \( S_{X_2} \) embedded into a common space (\( \mathbb{C}^n \) in our case), the set
\[
S_{X_1} + S_{X_2} = \{x_r + x_i : x_r \in S_{X_1}, x_i \in S_{X_2}\},
\]
is termed as the Minkowski sum of \( S_{X_1} \) and \( S_{X_2} \). Figure 1 illustrates this operation for two complex coplanar sets [17]. If, additionally, the arguments of the sum belong to linear and finite dimensional subspaces whose pairwise intersection is the origin, the operation is termed as Minkowski direct sum of the sets and is represented by \( S_{X_1} \oplus S_{X_2} \). Figure 2 illustrates this operation between two complex one-dimensional sets. The complete plot would require four real dimensions, but due to the limitations of the axonometric representation we only plot three of them.

### III. BOUNDED COMPONENT ANALYSIS ASSUMPTIONS

The bounded component analysis of the observations is based on two assumptions that concern the geometry of the convex support of the observations:

**A1.** (Compactness) The convex support of the target component \( S_{X_1} \), is compact (bounded and closed).

**A2.** (Direct decomposition) The convex support of the observations \( S_X \) is a Minkowski direct sum of the convex supports of the component signals, i.e.
\[
S_X = S_{X_1} \oplus S_{X_2}.
\]
Assumption A1 puts the emphasis on the relevance of the boundary of the target component, while assumption A2 describes the required geometric structure for the convex support set of the observations $S_X$. We show in the next subsection that the considered geometric structure should satisfy two conditions, according to the Minkowski direct sum in (13). The first one is that the convex support of the observations $S_X$ should expand the greatest possible set which is compatible with $S_X$, and $S_{X'}$, the convex supports of the components. The second one is that the components should belong to complementary subspaces, which is consistent with our initial assumption in equation (2).

A. The implications of the direct decomposition

The next lemma, whose proof is given in appendix A, tries to clarify the geometric meaning and implications of the bounded component analysis assumption A2.

**Lemma 1:** There is an equivalence between assumption A2 and the following set of necessary and sufficient conditions:

A2-a. (Factorization) The joint convex support of the components decomposes as the Cartesian product of their marginal convex supports, i.e.,

$$ S_{(X_{xt}, X_{xv})} = S_{X_{xt}} \times S_{X_{xv}}. \quad (14) $$

A2-b. (Uniqueness of the representation) The components $X_{xt}$ and $X_{xv}$ belong to complementary subspaces, i.e., $\langle X_{xt} \rangle \cap \langle X_{xv} \rangle = \{0\}$. Thus, for each possible observation $x \in S_X$ there is a unique decomposition of the form $x = x_{xt} + x_{xv}$, where $x_{xt} \in S_{X_{xt}} \subset \langle X_{xt} \rangle$ and $x_{xv} \in S_{X_{xv}} \subset \langle X_{xv} \rangle$.

B. Ties with Independent Component Analysis

The following lemmas explore the strong links between independent component analysis and bounded component analysis.

**Lemma 2:** Condition A2-a is weaker than the mutual independence of the components.

**Proof:** When $X_{xt}$ and $X_{xv}$ are independent, their joint density decomposes as the product of the marginals

$$ p(x_{xt}, x_{xv}) = p_{X_{xt}}(x_{xt})p_{X_{xv}}(x_{xv}), \quad (15) $$

and its convex support can be written as the Cartesian product of the convex supports of the marginals: $S_{(X_{xt}, X_{xv})} = S_{X_{xt}} \times S_{X_{xv}}$. Therefore, condition A2-a holds true when the components are independent.

**Corollary 1:** The mutual independence of the components together with their restriction to belong to complementary subspaces (property A2-b) implies the direct decomposition assumption (A2) of BCA.

**Proof:** This corollary follows from the combination of lemmas 1 and 2. However, one may find the next argument more instructive. For mutually independent components, the density of the observations equals the convolution of the marginal densities

$$ p_X(x) = \int \delta(x - (x_{xt} + x_{xv})) p(x_{xt}, x_{xv}) dx_{xt} dx_{xv}, \quad (16) $$

$$ = \int p_{X_{xt}}(x_{xt}) p_{X_{xv}}(x - x_{xt}) dx_{xt}, \quad (17) $$

$$ = p_{X_{xt}}(x) * p_{X_{xv}}(x). \quad (18) $$

Equating the support of both sides of the expression one obtains the decomposition as a Minkowski sum

$$ S_X = S_{X_{xt}} + S_{X_{xv}}. \quad (19) $$

Additionally, if the components belong to complementary subspaces, then this is also a Minkowski direct sum, yielding assumption A2.

IV. INNER PARAMETRIZATION OF THE COMPONENTS

It is not practical to explore directly the $n$-dimensional space of the observations for trying to find a target component, which is usually of smaller dimensionality ($p \ll n$). Instead, it is convenient to define an inner parametrization of this $p$-dimensional complex subspace. This is implicitly done by

$$ X_{xt} = A_\ast S_{xt}, \quad (20) $$

where $S_{xt} \in C^p$ denotes target source (the coordinates of this latent component relative to the $p$-dimensional subspace where it lies) and the full column-rank matrix $A_\ast \in C^{n \times p}$, whose columns determine the subspace $\langle X_{xt} \rangle$ of the component. However, there is an intrinsic ambiguity in the inner parametrization of a component even when its subspace is fixed.

**Proposition 1:** (Essential equivalence) The subspace of the target component $\langle X_{xt} \rangle$ determines the tuple (target source, target mixing system) up to a set of invertible linear transformations, in the sense, that for any non-singular matrix $M \in C^{n \times n}$ the following pairs of basis and coordinates are essentially equivalent

$$ (A_\ast, S_{xt}) \equiv (A_\ast M, M^{-1} S_{xt}). \quad (21) $$

This result can be seen as a generalization of the familiar scaling and phase indeterminacy (between the sources and the mixing columns) for one-dimensional components [1]. Its proof is straightforward because it is consequence of the fact that any of both pairs generates the same component $X_{xt} = A_\ast S_{xt} = (A_\ast M)(M^{-1} S_{xt})$.

Analogously to (20), we can parameterize the complementary component as $X_{xv} = A_{xv} S_{xv}$, where the full column-rank matrix $A_{xv} \in C^{n \times n-p}$ spans the complementary part of the signal space of the observation that was not generated by $A_\ast$.

**Note that** $A_{xt}$ and $A_{xv}$ determine the algebraic structure of the components (their linear subspaces) while the random vectors $S_{xt}$ and $S_{xv}$ determine their corresponding statistical structure.

Our previous definitions lead to the familiar linear mixing model for the observations random vector

$$ X = A S, \quad (22) $$

where $A = (A_{xt}, A_{xv}) \in C^{n \times n}$ is the mixing matrix and $S = (S_{xt}, S_{xv})^T \in C^n$ is the vector of sources.
The linear estimate of the sources $Y = \hat{S}$ can be obtained by pre-multiplying the observations by the non-singular matrix $B = A^{-1}$, which is an estimate of the inverse of the mixing system $A$. In this way, we define the output random vector

$$Y = \begin{pmatrix} Y_x \\ Y_\nu \end{pmatrix} = BX,$$

where $Y_x \equiv \hat{S}_x$ and $Y_\nu \equiv \hat{S}_\nu$.

We are mainly interested in the extraction of the target source, so only the estimate

$$Y_\tau = B_\tau X$$

needs to be evaluated in practice. The random vectors $Y$ and $Y_\nu$ are only auxiliary instruments for the theoretical justification of the extraction criteria. However, their estimation will not be required in the evaluation of the criteria for the extraction of the target source.

A. The implications of the BCA assumptions for $S$ and $A$

Once the mixing matrix $A$ in (22) is fixed, there is the bijective correspondence between the observations and the sources

$$X = AS \leftrightarrow S = A^{-1}X.$$  

(25)

Let the notation $0_k \equiv (0, \cdots, 0)^T$ denote zero vector of dimension $k$. Equation (25) also establishes a correspondence between the direct decomposition of $S_X$ and the factorization of $S_S$ as a Cartesian product of the convex supports of the target and nuisance components, i.e.,

$$S_X = S_{x_1} \otimes S_{x_2} \otimes S_{x_3} \leftrightarrow S_S = S(g_{x_1}) \otimes S(g_{x_2}),$$

(26)

$$= S_{x_1} \times S_{x_2}.$$  

(27)

This reveals a direct mapping of the BCA assumptions for the observations ($A_1, A_2-a$ and $A_2-b$) into an equivalent set of hypotheses ($H_1$, $H_2$ and $H_3$) for the sources and the mixing matrix:

H1. (Compactness) The convex support of the desired source $S_{x_1} \subset \mathbb{C}^p$ is compact.

H2. (Factorization) The convex support of the joint vector of sources decomposes as the Cartesian product of the convex supports of its marginal components, i.e.,

$$S_S = S_{x_1} \times S_{x_2}.$$  

(28)

H3. (Lossless mixing) The mixing system is characterized by a non-singular mixing matrix $A \in \mathbb{C}^{n \times n}$.

One can observe the mapping between assumptions and hypothesis in the correspondence of: $A_1$ with $H_1$, $A_2-a$ with $H_2$, and $A_2-b$ with $H_3$.

V. ESTIMATING THE DENSITY OF THE OBSERVATIONS

Even in absence of a model for the observed data $X^{1:T}$, one can build the empirical density estimate

$$p_X(x) = \frac{1}{T} \sum_{t=1}^{T} \delta(x - x(t)), \quad \delta(\cdot) \text{ denotes the Dirac delta function.}$$

A nice property of this non-parametric estimate is that the statistics of a random variable $X^T$—with density $p_X(x)$—match with the sample statistics of the observations.

However, when some hypotheses or assumptions are available about the structure of the observations, a semiparametric statistical model can be a convenient way to take this extra information into account. Let the operator $| \cdot |$ denote the modulo of the determinant of the matrix argument. From the model of the outputs in (23), one can solve for the observations as $X = B^{-1}Y$ and use the Jacobian determinant $|BB^H|$ of this change of variables to write the density of the observations in terms of the density of the outputs

$$p_X(x) = |BB^H| p_Y(y).$$

(30)

In practice, the density of the outputs is unknown, so we propose to replace it by the factorized estimate $\hat{p}(y) = \hat{p}_{Y_\tau}(y_\tau)\hat{p}_{Y_\nu | Y_\tau}(y_\nu | y_\tau)$. This way, we obtain the semiparametric estimate of the density of the observations

$$\hat{p}_X(x) \equiv \hat{p}_X(x; \theta) = |BB^H| \hat{p}_{Y_\tau}(y_\tau)\hat{p}_{Y_\nu | Y_\tau}(y_\nu | y_\tau), \quad \theta = (B, \hat{p}_{Y_\tau}, \hat{p}_{Y_\nu | Y_\tau}),$$

(31)

where the tuple

$$\theta = (B, \hat{p}_{Y_\tau}, \hat{p}_{Y_\nu | Y_\tau}),$$

(32)

collects all the parameters and functional choices that determine (31).

Considering the BCA hypotheses and the evidence of the observed data, we now face the problem of choosing the elements in $\theta$ that provide the robust fit of $\hat{p}_X(x)$ to the unknown density of the observations. For this purpose, we propose to work with a group of summary statistics of the sample set of outputs

$$y^{1:T} = \begin{pmatrix} y^{1:T}_\tau \\ y^{1:T}_\nu \end{pmatrix} = B \begin{pmatrix} x^{1:T} \end{pmatrix}. $$

(33)

On one hand, the natural key statistic in BCA for the target component is the convex support of the target output

$$S_{Y_\tau} = \text{cl conv } y^{1:T}_\tau.$$  

(34)

On the other hand, let us denote the sample covariance matrix of the target output $y^{1:T}_\tau$ by $\Sigma_{y_\tau}$, and the sample cross-covariance matrix between the target and nuisance outputs by $\Sigma_{y_\tau, y_\nu}$, where both are biased and consistent estimates. One can obtain relevant information about the nuisance component from the first and second order conditional statistics of the nuisance outputs given the target outputs:

$$\mu_{y_\tau | } = \Sigma_{y_\nu, y_\tau}\Sigma_{y_\nu}^{-1} y_\tau, \quad \Sigma_{y_\nu, y_\tau} = \Sigma_{y_\nu} - \Sigma_{y_\nu, y_\tau}(\Sigma_{y_\nu})^{-1}\Sigma_{y_\nu, y_\tau}.$$  

(35)

and

(36)

After grouping the partial evidence extracted from the output data in the set of sample statistics

$$\pi_y = \left\{ S_{Y_\tau}, \mu_{y_\tau | y_\nu}, \Sigma_{y_\nu | y_\tau} \right\}, $$

(37)

Semiparametric statistical models has been used in [37] to describe and analyze the blind source separation problem.
we define the set of admissible density functions for the observations as
\[
P_X(B, \pi_Y) = \{ \hat{p}_X(x) : \text{eq. (31) holds true for } S_{\pi_Y} \subseteq S_Y, \\
\mu_{Y \mid x} = \mu_{Y \mid \hat{p}_X}, \Sigma_{Y \mid x} = \Sigma_{Y \mid \hat{p}_X} \},
\]
which is the closed and convex\(^3\) set of distributions of the observations \(P_X \subseteq P\) that we consider compatible with our signal model and with the partial evidence collected in \(\pi_Y\).

Hereafter, unless stated otherwise, we will assume in our analyses that the following condition holds true.

**Regularity condition 1:** The density of the observations belongs to the admissible set, i.e.,
\[
p_X \in P_X(B, \pi_Y).
\]

This condition is asymptotically satisfied for large samples when not only the target source is bounded, but also is the nuisance source. The boundedness assumption guarantees the convergence, for sufficient large \(T\), of the sample convex support to the true convex support \((S_{\pi_Y} \to S_Y)\). On the other hand, the weak law of large numbers [38, pag. 65] guarantees that asymptotically the sample averages converge in probability to the true averages (so \(\mu_{Y \mid x} \to \mu_{Y \mid \hat{p}_X}\) and \(\Sigma_{Y \mid x} \to \Sigma_{Y \mid \hat{p}_X}\).

For the same reason, the normalized log-likelihood of the set of the observations converges in probability to the expected log-likelihood of the observations vector
\[
l(\theta) = \frac{1}{T} \log \hat{p}_X^{1:T}(x^{1:T})
\]
\[
= \frac{1}{T} \sum_{t=1}^{T} \log \hat{p}_X(x(t)) \xrightarrow{p} \int p(x) \log \hat{p}_X(x) dx,
\]
which can be used to measure the relative goodness of fit [39], [30]. Its opposite is the complexity loss function
\[
\psi(p_X, \hat{p}_X) = \int p_X(x) \log \frac{1}{\hat{p}_X(x)} dx
\]
that measures the expected bits required for describing observations coming from the true density \(p_X(x)\) with a coding scheme based on the estimate \(\hat{p}_X(x)\) (see [25, pag. 466]).

\section{Minimax complexity criterion}

The minimax criterion gives an insurance against the worst case situation because it aims to minimize the expected complexity in the least favorable case [26]. We can define a risk for each estimate \(\hat{p}_X(x) \in P\) that represents a valid distribution. The risk function
\[
R_\psi(\hat{p}_X) = \sup_{p_X \in P_X(B, \pi_Y)} \psi(p_X, \hat{p}_X),
\]
measures the worst case expected complexity of each observation over the set of admissible distributions \(P_X\).

On the basis of the given information, when minimizing the risk one obtains a parsimonious and robust description of the observations through the density estimate
\[
\hat{p}_X^*(x) = \arg \inf_{\hat{p}_X \in P_X(B, \pi_Y)} R_\psi(\hat{p}_X)
\]
\[
= \arg \inf_{\hat{p}_X \in P_X(B, \pi_Y)} \sup_{p_X \in P_X(B, \pi_Y)} \psi(p_X, \hat{p}_X).
\]

The solution of this class of optimization problems was studied by Topsøe [27] and Kullback [28] (among many others), and it is known to lead to the principle of maximum entropy estimation of Jaynes [29]. The robustness lemma [27] introduces the solution as a robust estimate of the density of the observations and presents its equilibrium property for the encoding complexity. These results are reformulated in the next theorem.

**Theorem 1 (Robustness and equilibrium):** When the distribution of maximum entropy within the admissible set \(P_X(B, \pi_Y)\) exists, it yields \(\hat{p}_X^*\), the best robust estimate of \(p_X\) in the sense of attaining the minimum risk \(R_\psi(\hat{p}_X^*)\) in the least favorable case. Moreover, for a coding scheme based on \(\hat{p}_X^*\), the encoding complexity of any of the members of the admissible set is constant (or in equilibrium)
\[
R_\psi(\hat{p}_X^*) = \psi(p_X, \hat{p}_X^*), \quad \forall p_X \in P_X(B, \pi_Y).
\]

We have seen in (45) that the minimum risk solution depends on the unknown density of the observations, thus, we face the problem of how to evaluate \(R_\psi(\hat{p}_X^*)\) when only a set of observations is given. The next corollary provides a practical answer to this question.

**Corollary 2 (Bridging the evaluation gap):** Under the regularity condition 1, one can exactly obtain the minimum risk solution from the samples just replacing the unknown distribution \(p_X\) by its empirical estimate \(\hat{p}_X\), i.e.,
\[
R_\psi(\hat{p}_X^*) = \psi(\hat{p}_X^*, \hat{p}_X) = -\frac{1}{T} \sum_{t=1}^{T} \log \hat{p}_X^*(x(t)) \equiv l(\theta^*).
\]
where \(l(\theta^*)\) denotes the normalized log-likelihood of the observations for the robust density model \(\hat{p}_X^*(x)\).

**Proof:** The proof of the corollary relies on the equilibrium property of the encoding complexity over the admissible set \(P_X(B, \pi_Y)\). This property ensures the equality \(\psi(p_X, \hat{p}_X^*) = \psi(p_X^O, \hat{p}_X^*)\) because the regularity condition is ensuring that \(p_X \in P_X(B, \pi_Y)\), while, by construction, the empirical density satisfies that \(p_X^O \in P_X(B, \pi_Y)\). \(\square\)

On the other hand, when the regularity condition 1 is not true, we should expect an approximation error. However, for sufficient large samples we still have the asymptotic convergence \(\psi(p_X^O, \hat{p}_X^*) \xrightarrow{p} \psi(p_X, \hat{p}_X^*)\) as a consequence of (40).

**Lemma 3 (Minimum risk distribution):** For the admissible set of densities defined in (38), the distribution of minimum risk is given by
\[
\hat{p}_X^*(x) = \left[ BB \right]_{\psi}^{\psi} (p_{\psi Y}(y^r) p_{Y \mid Y^r} (y^r | y^r)),
\]
where
\[
p_{\psi Y}(y^r) = \begin{cases} \frac{1}{|\mathcal{Y}^r| (S_{\pi_Y})} & \text{for } y^r \in S_{\pi_Y}, \\ 0 & \text{for } y^r \not\in S_{\pi_Y}, \end{cases}
\]
and
\[ p_{Y^N|Y}(y_v|y_r) = \frac{e^{-(y_v - \mu_{y_v}(y_r))^T \Sigma_{y_v}^{-1}(y_v - \mu_{y_v}(y_r))}}{\pi^{n_p} |\Sigma_{y_v}|}. \] (49)

See the Appendix C for the proof.

In order to rewrite this solution into a more convenient expression for its interpretation, we need to introduce some additional notation. Consider a Gaussian random vector \( X^N \), which shares the sample mean \( \mu_x \) and the sample covariance matrix \( \Sigma_x \) of the data \( x^{1:T} \). The density of \( X^N \) can be regarded as a robust approximation to the density of the observations.

**Lemma 4 (The Gaussian approximation):** Consider the family \( \mathcal{P}_x^N(\{\mu_x, \Sigma_x\}) \) of all the densities with mean \( \mu_x \) and covariance matrix \( \Sigma_x \). The robust approximation to \( p_X(x) \) within this family is the Gaussian distribution
\[ p_{X^N}(x) = \arg \inf_{p_X \in \mathcal{P}_x} \sup_{p_{Y^N} \in \mathcal{P}_y^N(\{\mu_Y, \Sigma_Y\})} \psi(p_X, p_{Y^N}). \] (50)

The lemma is a direct consequence of the fact that the Gaussian distribution maximizes the entropy among all distributions with a given mean and covariance. Thus, from theorem 1 the density \( p_{X^N}(x) \) yields the robust approximation to \( p_X(x) \) in the sense of (50).

Similarly, the robust approximation to the density \( p_{Y^N}(y_r) \) on the basis of its sample mean \( \mu_Y \) and sample covariance \( \Sigma_Y \) is given by \( p_{Y^N}(y_r) \) where \( Y^N = BX^N \). Lastly, we obtained in (48) the uniform density \( p_{Y^N}(y_r) \) as robust approximation of \( p_{Y^N}(y_r) \) among all densities whose convex support is contained within \( S_Y \), the convex support of the output samples. With these definitions, we are ready for the formulation of the next lemma.

**Lemma 5 (Reinterpretation in factorized form):** The density that minimizes the risk in lemma 3 can be factorized as
\[ \hat{p}_X(x) = p_{X^N}(x) \frac{p_{Y^N}(y_r)}{p_{Y^N}(y_r)}. \] (51)

allowing us to reinterpret \( \hat{p}_X(x) \) as a correction of the robust Gaussian approximation of \( p_X(x) \) by a factor that involves the robust uniform and Gaussian density approximations of \( p_{Y^N}(y_r) \).

**Proof:** Consider the transformation of random vectors \( Y^N = BX^N \) and the associated transformation rule for the involved distributions
\[ |BB^T| = \frac{p_{X^N}(x)}{p_{Y^N}(y_r)}. \] (52)

By substituting (52) in equation (47) and performing a straightforward cancelation of common terms, the factorized result of this lemma is obtained. The interpretation of equation (51) in the last part of the lemma easily follows from the definitions of the involved factors.

\[ \square \]

**B. Kullback-Leibler divergence and minimum risk**

We have seen that the density estimate of the observations \( \hat{p}_X(x) \) depends on the value of the separation matrix \( B \) and on the functional choices of \( \hat{p}_{Y_r} \) and \( \hat{p}_{Y_r|Y_r} \). In the previous subsection, we determined the functional part of the estimate that minimizes the risk \( R_\psi(\hat{p}_X) \), resulting the robust density estimate \( \hat{p}_X(x) \) given in (51). However, this result still depends on our choice of the extraction matrix \( B \). For this reason, hereafter, the minimum risk \( R_\psi(\hat{p}_X) \) will be considered, as the underlying risk to minimize with respect to the extraction matrix.

This subsection relates the minimum risk criterion with the Kullback-Leibler divergence of the robust distribution \( \hat{p}_X(x) \) from the distribution \( p_X(x) \). Since these densities satisfy the condition that \( \hat{p}_X(x) = 0 \Rightarrow p_X(x) = 0 \), their Kullback-Leibler (KL) divergence is well defined and given by
\[ D(p_X||\hat{p}_X) = \int p_X(x) \log \frac{p_X(x)}{\hat{p}_X(x)} \, dx. \] (53)

A simple manipulation of this expression reveals its decomposition in terms of the entropy of the observations and the minimum risk, i.e.,
\[ D(p_X||\hat{p}_X) = R_\psi(\hat{p}_X) - h(p_X). \] (54)

Since the entropy of the observations is independent of the extraction matrix there is an equivalence between the optimization of the minimum risk and the fitting (in KL divergence sense) of the robust distribution \( \hat{p}_X(x) \) to the distribution \( p_X(x) \).

**C. Negentropy-likelihood criterion and minimum risk**

The minimum risk criterion can also be linked with a version of the negentropy-likelihood maximization criterion for the recovery of a target source. This negentropy-likelihood criterion was defined in the context of independent component analysis (see [31, pag. 866]) as the difference of Kullback-Leibler divergences
\[ Q_{EL}(Y_r, \hat{p}_{S_r}) = D(p_{Y_r} || p_{Y^N}) - D(p_{Y_r} || \hat{p}_{S_r}). \] (55)

It is a combination of two unsupervised extraction criteria, which avoids the difficulty of resorting to an accurate estimation and optimization of the true entropy of the target output \( h(p_{Y_r}) \). The criterion combines two forces that drive our search of the desired distribution. In the maximization of \( Q_{EL}(Y_r, \hat{p}_{S_r}) \), the negentropy part of the criterion \( D(p_{Y_r} || p_{Y^N}) \) drives \( p_{Y_r} \) as far as possible from the Gaussian distribution, while, at the same time, the likelihood part of the criterion \( -D(p_{Y_r} || \hat{p}_{S_r}) \) drives \( p_{Y_r} \) as close as possible to a given reference model \( \hat{p}_{S_r} \) for the density of the target source.

The next lemma presents a useful decomposition of the minimum risk criterion, and the equivalence of its minimization with the maximization of a negentropy-likelihood criterion that pursues a balance between the non-Gaussianity and uniformity of \( Y_r \).

**Lemma 6 (Decomposition of the minimum risk):** Under the regularity condition 1, the solution of minimum risk decomposes as
\[ R_\psi(\hat{p}_X) = \psi(p_X^0, p_{X^N}) - \Delta \psi(B_r). \] (56)

The first term
\[ \psi(p_X^0, p_{X^N}) = \log (\pi e)^n |\Sigma_x|, \] (57)
represents the complexity of the second order approximation to the density of the observations seen in lemma 4, and is independent of $\mathbf{B}_\tau$. On the other hand, the last term

$$\Delta \psi(\mathbf{B}_\tau) = \psi(p_{\mathbf{Y}^U} | p_{\mathbf{Y}'}) - \psi(p_{\mathbf{Y}^U} | p_{\mathbf{Y}'})$$

(58)

is a maximization criterion. It measures the expected number of bits that we save in the encoding of the outputs when they are modeled with a uniform distribution rather than with a Gaussian. This difference of encoding complexity is indeed a negentropy-likelihood criterion

$$\Delta \psi(\mathbf{B}_\tau) \equiv Q_{EL}(\mathbf{Y}_\tau, p_{\mathbf{Y}'})$$

(59)

with a reference model based on $p_{\mathbf{Y}'}$, the uniform distribution of equation (48) that best fitted the considered hypothesis and the partial evidence in (37).

The proof of the lemma is presented in Appendix D.

Note that the negentropy-likelihood criterion was a fitting criterion of the density of the observations that assumed the considered reference model for the density of the desired source as the true one. However, the lemma implies that this reference does no longer need to match the truth, because of its reinterpretation as a robust approximation that minimizes the minimum risk. This interpretation is clarified by the following remark.

Remark 1 (The target source doesn’t need to be uniform): The fact that the minimum risk solution lead to the maximization of the negentropy-likelihood criterion with a reference model for the target source based on $p_{\mathbf{Y}'}$, does not imply that we are assuming the target source to be uniform. On the contrary, it was the bounded hypothesis for the target source what has led to the choice of the support of the output as sufficient statistic for this assumption, and this eventually has driven our best strategy for minimizing the risk towards the optimization of the negentropy-likelihood criterion with a reference model based on the uniform distribution.

D. The minimum normalized volume criterion

This subsection justifies the proposal of the minimum normalized volume as a natural minimum risk criterion for the bounded component analysis hypothesis.

In the context of BCA, the minimization of the normalized volume of the target output yields a parsimonious criterion that aims to obtain the most compact explanation for the observations, in the sense of attaining the minimum volumetric extension for its convex support. The normalized volume can be defined as

$$N_{Vol}(\mathbf{B}_\tau) = \frac{V_{C^\circ}(S_{\mathbf{Y}'})}{|\Sigma_{\mathbf{Y}'_\tau}|}$$

(60)

where $V_{C^\circ}(S_{\mathbf{Y}'})$ denotes the volume of the convex support of the uniform random vector $\mathbf{Y}'$ that best matches with the given hypothesis and with the considered evidence on the support of the data. Under the transformation of the outputs

$$\mathbf{Y}'_\tau = \Sigma_{\mathbf{Y}'_\tau}^{-1/2} \mathbf{Y}'_\tau$$

(61)

which normalizes the dispersion of $\mathbf{Y}'_\tau$, we have the volume transformation $V_{C^\circ}(S_{\mathbf{Y}'}) = |\Sigma_{\mathbf{Y}'_\tau}|^{-1} V_{C^\circ}(S_{\mathbf{Y}'})$. Thus, the criterion is also compactly rewritten as

$$N_{Vol}(\mathbf{B}_\tau) = V_{C^\circ}(S_{\mathbf{Y}'})$$

(62)

Note that there will be no loss of generality in considering the logarithmic version of the criterion log $N_{Vol}(\mathbf{B}_\tau)$, because the arguments of the extrema of a function are keep invariant under strictly monotonous ascending transformations.

Under the BCA hypothesis H1-H3 and the evidence of $S_{\mathbf{Y}'_\tau}$ considered in (37), the support of the robust uniform approximation is $S_{\mathbf{Y}'_\tau} = S_{\mathbf{Y}_\tau}$, so the logarithmic version of the normalized volume simplifies to

$$C(\mathbf{B}_\tau) = \log \frac{Vol_{C^\circ}(\mathbf{Y}_\tau)}{|\Sigma_{\mathbf{Y}'_\tau}|} = \log Vol_{C^\circ}(\hat{\mathbf{y}}_\tau),$$

(63)

where $\hat{\mathbf{y}}_\tau = \Sigma_{\mathbf{Y}'_\tau}^{-1/2} \mathbf{y}_\tau$ refers to the standardized samples of the target output.

The next theorem aims to summarize most of the previously obtained results.

Theorem 2 (The interpretation of the BCA solutions): In bounded component analysis, under the regularity condition 1, there is an equivalence between the criteria of:

a) minimum complexity risk

$$\mathbf{B}_\tau^* = \arg \min_{\mathbf{B}_\tau} R_\psi(\mathbf{Y}_\tau),$$

(65)

b) max. likelihood of the observations based on $\mathbf{Y}_\tau(x)$

$$\mathbf{B}_\tau^* = \arg \max_{\mathbf{B}_\tau} \frac{1}{T} \sum_{t=1}^{T} \log p_\mathbf{Y}_\tau(x(t)),$$

(66)

c) minimum KL diverge of the estimate $\hat{\mathbf{Y}}_\tau(x)$ from the distribution of the observations

$$\mathbf{B}_\tau^* = \arg \min_{\mathbf{B}_\tau} D(p_\mathbf{Y} || \hat{p}_\mathbf{Y}_\tau),$$

(67)

d) maximum negentropy-likelihood of the target output with a reference model based on the uniform distribution

$$\mathbf{B}_\tau^* = \arg \max_{\mathbf{B}_\tau} V_{C^\circ}(S_{\mathbf{Y}'})$$

(68)

e) minimum normalized volume of the target output

$$\mathbf{B}_\tau^* = \arg \min_{\mathbf{B}_\tau} C(\mathbf{B}_\tau).$$

(69)

All these criteria yield a wealth of complementary interpretations for the solutions to the problem of bounded component analysis.
The theorem is a direct consequence of the relations between these criteria, which have been presented in the equations: (46), (54), (56) and (64).

VI. SHAPE HYPOTHESES FOR THE TARGET SOURCE

In most applications, it is reasonable to expect that the convex support of the target p-dimensional source \( S = (S_1, \ldots, S_p)^T \) can be further decomposed in terms of the convex support sets of its components, satisfying the following hypothesis:

**H4. (Factorization of the target component)** The convex support of the target multidimensional source decomposes as the Cartesian product of the convex supports of its one-dimensional components, i.e.,

\[
S_S = S_{S_1} \times \cdots \times S_{S_p}. \tag{70}
\]

Let \( a_i \) denote the \( i^{th} \)-column of \( A_\tau \). With the additional assumption, the support of the observations decomposes in terms of the one-dimensional components \( X_i = a_i S_i, \ i = 1, \ldots, p \) and of the nuisance component \( X_y \), as

\[
S_X = (S_{X_1} + \cdots + S_{X_p}) \cup S_{X_y}. \tag{71}
\]

The hypothesis H4 can be incorporated to our previous analysis where we fitted the estimate \( \hat{p}_X(x) \) to the density of the observations. Due to the decomposition of the convex support of the target source \( S_S \), it is enough to gather the evidence about the convex support of its component variables \( S_{S_1}, \ldots, S_{S_p} \). Thus, a new admissible set \( \mathcal{P}_X(B, \pi^2) \) for the density of the observations is built on the basis of the following partial evidence from the outputs

\[
\pi^1 = \{ S_{y_1}, \ldots, S_{y_p}, \mu_{y_\tau|y}, \Sigma_{y_\tau|y} \}, \tag{72}
\]

where \( y = (y_1, \ldots, y_p)^T \) and \( y = (y_1^T, y_p^T)^T \). Under the hypotheses H1-H4, following the same steps as before, the estimated density that minimizes the risk can be obtained as

\[
\hat{p}_X^2(x) = p_{X^n}(x) \prod_{i=1}^P p_{Y^n}(y_i) / p_{Y^n}(y_T), \tag{73}
\]

where each \( p_{Y^n}(y_i) \) denotes a uniform distribution of support \( S_{y_i} \). Minimizing the risk maximizes the likelihood of the observations with a model based on \( \hat{p}_X^1(x) \), while it minimizes the KL fit of the density estimate \( \hat{p}_X^2(x) \) to \( p_X(x) \). It also maximizes the negentropy-likelihood criterion

\[
\Delta \psi^1(B_\tau) = D(p_Y || p_{Y^n}) - \sum_{i=1}^P D(p_{Y_i} || p_{Y^n}), \tag{74}
\]

and minimizes a different version of the minimum normalized volume criterion. Let

\[
A(y_i) = Vol_{C^1}(y_i) \equiv Vol_{H^2}(\Re\{y_i\}, \Im\{y_i\}) \tag{75}
\]

denote the area measure of the convex hull for each set of complex outputs \( y_i^{1T}, \ i = 1, \ldots, p \). The new criterion for the extraction of a subset of \( p \) complex bounded sources is now given by

\[
C^1(B_\tau) = \log \left( \prod_{i=1}^P A(y_i) / |\Sigma_{y\tau}| \right). \tag{76}
\]

The assumption H4 is only one of the many hypotheses that could be considered for the shape of the target multidimensional source. So the presented framework is quite general. However, instead of assuming alternative geometries for the multidimensional target component, we consider more illustrative to present our argument with a specific shape assumption for the one dimensional components \( S_{1}, \ldots, S_{p} \). For instance, we can define the geometry of the convex support of the target sources as part of our hypotheses:

**H5. (Example: a generalized disk shape hypothesis)** The shapes of the convex supports \( S_{S_i} \) of the one-dimensional target sources take the form of \( q_i \)-disks with \( q_i \geq 1 \), i.e., for \( i = 1, \ldots, p \),

\[
D_{q_i}(c_i, r_i) \equiv \{ s_i \in C^1 : ||\{ \Re\{y_i(t) - c_i\}, \Im\{y_i(t) - c_i\} \} ||_2 \leq r_i \}, \tag{77}
\]

where \( D_{q_i}(c_i, r_i) \) denotes a complex \( q_i \)-disk of center \( c_i \) and radius \( r_i \).

The admissible set \( \mathcal{P}_X(B, \pi^2) \) for the density of the observations exploits the new partial evidence from the outputs

\[
\pi^2 = \left\{ c, r_1(y_1), \ldots, r_p(y_p), \mu_{y_\tau|y}, \Sigma_{y_\tau|y} \right\}, \tag{78}
\]

where the vector \( c = (c_1, \ldots, c_p)^T \) collects the centers of the disks, while their radii for \( i = 1, \ldots, p \) are defined as

\[
r_i(y_i) = \max_{t_i=1,\ldots,t} ||\{ \Re\{y_i(t) - c_i\}, \Im\{y_i(t) - c_i\} \} ||_2. \tag{79}
\]

If the center of each disk \( c_i \) is a priori known, the radius \( r_i \) is directly evaluated from (79). Otherwise, the parameters \( (c_i, r_i) \) of each disk are jointly determined so as to yield the \( q_i \)-disk of minimum area which encloses \( y_i(t) \). On the basis of the assumptions H1-H5, the density estimate that minimizes the risk is

\[
\hat{p}_X^2(x) = p_{X^n}(x) \prod_{i=1}^P p_{Y^n}(y_i) / p_{Y^n}(y_T), \tag{80}
\]

where each \( p_{Y^n}(y_i) \) denotes a uniform distribution of support \( D_{q_i}(c_i, r_i) \). The resulting BCA criterion consists in the minimization of

\[
C^2(B_\tau, (q_1, \ldots, q_p)) = \log \left( \prod_{i=1}^P V_C(D_{q_i}(c_i, r_i)) / |\Sigma_{y\tau}| \right). \tag{81}
\]

The volume is invariant under translation, so we can write

\[
V_C(D_{q_i}(c_i, r_i)) = V_C(D_{q_i}(0, r_i)) = r_i^q \cdot v_{q_i}, \tag{82}
\]

where \( v_{q_i} \) denotes the volume of a unit \( q_i \)-disk [42]

\[
v_{q_i} \equiv V_C(D_{q_i}(0, 1)) = (2 \pi / \Gamma(q_i + 1))^q / |\Sigma_{y\tau}|. \tag{83}
\]

Therefore, the criterion in (81) is simplified to

\[
C^2(B_\tau, (q_1, \ldots, q_p)) = \log \left( \prod_{i=1}^P v_{q_i} r_i^q / |\Sigma_{y\tau}| \right). \tag{84}
\]

Note that the volumes \( v_{q_1}, \ldots, v_{q_p} \) are independent of \( B_\tau \) and, for this reason, they are not required for the optimization of the criterion. Although, they are needed for the purpose
of providing a consistent minimum risk comparison between criteria with alternative sets of hypotheses.

As one incorporates more information into the model, the resulting criteria are progressively more discriminative. For the sake of the completeness of our study, it is interesting to consider the following last hypothesis:

H6. (A common q-disk shape hypothesis) The convex supports of the target sources \( S_i \) for \( i = 1, \ldots, p \), share the common shape \( D_q(c, r) \).

For a given \( c \), the radius of the q-disks can be estimated by

\[
\tilde{r}(y_r) = \max_{t=1, \ldots, T} \max_{i=1, \ldots, p} \left( \| (R\{y_i(t) - c\}, \mathcal{D_q}(y_i(t) - c) \|_q \right). \quad (85)
\]

Otherwise, if \( c \) is unknown, one should jointly estimate the parameters of the q-disk \( D_q(c, r) \) of minimum area which encloses the samples of \( y_i(t) \) for \( t = 1, \ldots, T \), and \( i = 1, \ldots, p \). The partial evidence \( \pi_{y_r} = \{c, r(y_r), \mu_{y_r}, \Sigma_{y_r}\} \), together with the hypotheses H1-H6 define the new admissible set \( \mathcal{P}_3(Y, \pi_{y_r}) \). This yields the robust estimate

\[
\hat{p}^3_\mathcal{P}(x) = p_{\mathcal{P}^3}(x) = \prod_{i=1}^{p} \frac{p_{Y^p}(y_i)}{p_{Y^p}(y_r)} \cdot (86)
\]

where \( p_{Y^p}(y_i) \) denotes the uniform distribution with support on \( D_q(c, r(y_r)) \). Thus, its associated BCA criterion is given by

\[
C_3(B_r, q) = \log \left( \frac{\sqrt{q} \cdot r(y_r)}{|\Sigma_{y_r}|} \right)^{2p}. \quad (87)
\]

VII. COHERENCE AND ACCURACY OF THE HYPOTHESES

For the validity of the proposed method, each new assumption for the outputs should preserve the coherence with the previous ones. This coherence is guaranteed when the new assumptions do not decrease the size of the admissible sets of distributions in the sense that

\[
\mathcal{P}_\mathcal{X}(B, \pi_{y_r}) \subseteq \mathcal{P}_\mathcal{X}(B, \pi_{y_r}) \subseteq \mathcal{P}_\mathcal{X}(B, \pi_{y_r}) \subseteq \mathcal{P}_\mathcal{X}(B, \pi_{y_r}). \quad (88)
\]

Indeed, this inclusion is satisfied by the progressive incorporation of the compatible shape hypotheses H4, H5 and H6 to the basic ones in BCA (H1-H3). Under the regularity condition 1, the main consequence of the enlargement of the admissible set is that the minimum risk should not decrease with each new correct and coherent assumption, yielding the hierarchy

\[
C(B_r) \leq C_1(B_r) \leq C_2(B_r, (q, \ldots, q)) \leq C_3(B_r, q). \quad (89)
\]

Depending on whether the additional assumptions for the sources are correct we can roughly distinguish three situations:

1. Correct hypotheses: In this case the equality holds at (89) for the extraction solution, so the criterion is more discriminative as one incorporates additional valid hypothesis.

2. Approximately correct hypotheses: In this situation the risk can also increase at the extraction solution because of the inaccuracy of one or more hypotheses. However, the criterion can still work when the upper-bound it yields is sufficient tight.

3. Wrong hypotheses: When the new assumptions are incompatible with the previous ones, in the sense that they do not satisfy (88), or when these are too inaccurate approximations of the properties of the target sources, the criterion may no longer be discriminative.

VIII. THE UNIFYING FRAMEWORK FOR BCA CRITERIA

Table II presents a summary of the minimum risk interpretation of several bounded component analysis criteria. In this table, the hypotheses together with the evidence determine the estimated density of minimum risk and the associated minimum normalized volume criterion. Next to the hypotheses there is a label that distinguishes whether they are defined for the case of real or complex observations. Existing works that have some relation with the criteria are referenced in the last column.

The theoretical results we have obtained for complex observations \((x \in \mathbb{C}^n)\) also extend to the real case \((x \in \mathbb{R}^n)\), although, the resulting BCA criteria for real observations slightly differ as a consequence of the redefinition of the densities and their complexity, which are now given by

\[
\psi(p_{X^p}, p_{X^N}) = \frac{1}{2} \log (2\pi e)^n |\Sigma_x| \quad (90)
\]

\[
\psi(p_{y^p}, p_{y^N}) = \frac{1}{2} \log (2\pi e)^p. \quad (91)
\]

Thus, the constant of lemma 7 is now given by

\[
constant_1 = \frac{1}{2} \log (2\pi e)^n - p |\Sigma_x|. \quad (92)
\]

We can see in table II how the optimization of the minimum risk criterion provides a unified framework for most of the existing BCA criteria. The minimum range criterion was proposed by Pham in [9] for the blind separation of real and mutually independent sources of finite support. It was later extended to yield the simultaneous extraction of \( p \) sources [12] and reformulated as a bounded component analysis strategy [14]. This criterion

\[
C^R_1(B_r) = \log \left( \prod_{i=1}^{p} \frac{R(y_i)}{|\Sigma_{y_i}|^{1/2}} \right), \quad (93)
\]

is based on measurements of the statistical range of the outputs

\[
R(y_t) = \max_{i} \{y_i(t) - \min_i(y_i(t))\}, \quad i = 1, \ldots, p, \quad (94)
\]

and can be regarded as the first approach that pursues the minimization of a normalized volume of the outputs. Given the hypotheses H1-H4 in the real case, and the partial evidence \( \{c, R(y_r), \ldots, R(y_p), \mu_{y_r}, \Sigma_{y_r}\} \), the minimum risk principle provides a novel robust reinterpretation of (93).

When one adds the hypothesis of having sources of equal shape (H6 real case), the convex support of the \( p \)-dimensional target source is a hypercube. Under the hypotheses H1-H6 and the partial evidence \( \{c, Range(y_r), \ldots, \mu_{y_r}, \Sigma_{y_r}\} \), the minimum risk principle suggests a more discriminative version of the criterion

\[
C^R_3(B_r) = \log \left( \frac{Range(y_r)}{|\Sigma_{y_r}|^{1/2}} \right). \quad (96)
\]
Table II: Summary of the minimum risk interpretation of several BCA criteria for extraction ($p < n$) and separation ($p = n$).

<table>
<thead>
<tr>
<th>Hypotheses</th>
<th>Evidence Set</th>
<th>Robust Density Estimate</th>
<th>BCA Criterion</th>
<th>Ref.</th>
</tr>
</thead>
</table>
| H1-H3 complex | $\{S_{y_1}, \cdots, S_{y_p}, \mu_{y_i|y}, \Sigma_{y_i|y}\}$ | $\hat{p}_{x}^{C}(x) = p_{x|x}(x) \frac{p_{y|x}(y_{i})}{p_{y|x}(y_{r})}$ | $C(B_r) = \log \frac{\text{Vol}(C_r(y_r))}{|\Sigma_{y_i}|}$ | [
| H1-H3 real | equivalent real version | equivalent real version | $C^{B_r}(B_r) = \log \frac{\text{Vol}(B_r(y_r))}{|\Sigma_{y_i}|^{1/2}}$ | [9],[12] |
| H1-H4 complex | $\{S_{y_1}, \cdots, S_{y_p}, \mu_{y_i|y}, \Sigma_{y_i|y}\}$ | $\hat{p}_{x}^{C}(x) = p_{x|x}(x) \frac{\prod_{i=1}^{p} p_{y|x}(y_{i})}{p_{y|x}(y_{r})}$ | $C(B_r) = \log \frac{\prod_{i=1}^{p} A(y_{i})}{|\Sigma_{y_i}|}^{1/2}$ | [15],[16] |
| H1-H4 real | $\{c, R(y_1), \cdots, R(y_p), \mu_{y_i|y}, \Sigma_{y_i|y}\}$ | equivalent real version | $C^{B_r}(B_r) = \log \frac{\prod_{i=1}^{p} R(y_{i})}{|\Sigma_{y_i}|^{1/2}}$ | [9],[12] |
| H1-H5 complex | $\{c, r_{1}(y_1), \cdots, r_{p}(y_p), \mu_{y_i|y}, \Sigma_{y_i|y}\}$ | $\hat{p}_{x}^{C}(x) = p_{x|x}(x) \frac{\prod_{i=1}^{p} p_{y|x}(y_{i})}{p_{y|x}(y_{r})}$ | $C_{2}(B_r, q) \equiv \log \frac{\prod_{i=1}^{p} v_{i}r_{i}^{2}(y_{i})}{|\Sigma_{y_i}|}$ | [9],[18] |
| H1-H5 real | equivalent real version | equivalent real version | $C^{B_r}(B_r) \equiv \log \frac{\prod_{i=1}^{p} 2r(y_{i})}{|\Sigma_{y_i}|^{1/2}}$ | [9],[18] |
| H1-H6 complex | $\{c, r(y_r), \mu_{y_i|y}, \Sigma_{y_i|y}\}$ | $\hat{p}_{x}^{C}(x) = p_{x|x}(x) \frac{\prod_{i=1}^{p} p_{y|x}(y_{i})}{p_{y|x}(y_{r})}$ | $C_{2}(B_r, q) = \log \frac{(\sqrt{v_{q}r(y_{r})})^{2p}}{|\Sigma_{y_i}|}$ | [13] |
| H1-H6 real | $\{c, \text{Range}(y_r), \mu_{y_i|y}, \Sigma_{y_i|y}\}$ | equivalent real version | $C^{B_r}(B_r) = \log \frac{(\text{Range}(y_r))^{p}}{|\Sigma_{y_i}|^{1/2}}$ | |

The complex case adds greater freedom in the choice of the shape hypothesis for the target source, by diversifying the criteria. For instance, consider the different complex criteria $C_1(B_r)$ and $C_2(B_r(q_1, \ldots, q_p))$, $\forall q_i \geq 1$, $i = 1, \ldots, p$. We can observe in Table II that their real versions coincide, i.e. $C^{B_r}(B_r) = C^{B_r}(B_r)$, because the range and the radius for real one-dimensional sets satisfy $R(y_{i}) = 2r(y_{i})$. The simple geometric method proposed by Erdogan for the blind separation of complex sources [13] can also be reinterpreted as an implementation of the minimum risk principle. This geometric method proposes the minimization of the supremum of the $\infty$-norm of the real part of the standardized outputs, i.e.,

$$C_E(B_r) = \max \|\Re\{\hat{y}(t)\}\|_{\infty},$$

and it is based on the assumption of having standard sources with real and imaginary positive-negative peak symmetry, i.e., of complex sources of common square shape.

One can realize that this is the situation described by hypothesis H6 with $q = \infty$, center $c = 0$ and a replacement of estimate for the radius (half the size of the square) by

$$r'(y_r) = \max_{t} \max_{i} |\Re\{y_i(t)\}|.$$  (98)

Note that $r'(y_r)$ slightly differs from $r(y_r)$ in (85) because it omits the evidence coming from $3\{y(t)\}$. Under H1-H6 and the evidence summarized in $\{c, r'(y_r), \mu_{y_i|y}, \Sigma_{y_i|y}\}$, the minimization of the risk leads to a BCA strategy similar to $(87)$ with $\sqrt{v_{q}} = 2$ and $q = \infty$. For standardized outputs, it simplifies to

$$C_{2}(B_r, \infty) = \log (2r'(\hat{y}_r))^{2p} = \log (2C_E(B_r))^{2p},$$

which essentially coincides with (97), except for a strictly monotonic ascending transformation that leaves invariant the set of extremal arguments in both criteria.

IX. SIMULATIONS

In this section we intend to illustrate some of our theoretical claims with sample experiments. In order to ease the reproducibility of the results, the MatLab code used for the generation of Figures 3-6 is available in

http://personal.us.es/sergio/alg/risk.html

A. First experiment

Let $u_1(t), u_2(t) \in [0, 1]$ for $t = 1, \ldots, T$, denote two i.i.d. sequences of uniform samples with length $T = 500$. The $p$-dimensional target source $s_r(t) \in C^p$, with $p = 2$, is

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.
obtained by embedding these uniform samples into a higher dimensional space through the nonlinear transformation:

$$s_r(t) = \begin{pmatrix} s_1(t) \\ s_2(t) \end{pmatrix} = \begin{pmatrix} 2 (e^{2\pi u_1(t)})^2 \\ 2 (e^{2\pi u_2(t)})^2 \end{pmatrix}, \quad (100)$$

Although, we cannot plot the shape of this source due to its dimensionality, it is possible to show its projection ($\Re \{s_1(t)\}, \Im \{s_1(t)\}, \Re \{s_2(t)\}$) into the three dimensional real space, which adopts the cylinder-shaped form of figure 3(a).

Let us assume that $g_i(t), \ i = 1, \ldots, n-p$, denote the realizations of circular Gaussian i.i.d. complex processes that are mutually independent. The nuisance source $s_p(t) \in \mathbb{C}^{n-p}$ was set to

$$s_p(t) = \begin{pmatrix} g_1(t) \\ \vdots \\ g_{n-p}(t) \end{pmatrix} + \text{sign}(s_1(t) + s_2(t)) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad (101)$$

where $\text{sign}(\cdot)$ denotes the complex sign operator, i.e., $\text{sign}(x) = x/|x|$. The nuisance source is $(n-p)$-dimensional and is formed by the superposition of a circular Gaussian component and another component nonlinearly dependent on the target sources.

For greater generality, the simulations consider an over-determined problem with $m = 25$ sensors and $n = 15$ sources ($m > n$). This situation only requires a minor redefinition of the notation, because the projection of the observations onto the signal subspace is now assumed to be enforced by the internal implementation of the extraction system $B_r \in \mathbb{C}^{p \times m}$. We mix the sources with a complex random matrix $A \in \mathbb{C}^{m \times n}$ to obtain the observations $x(t) = As(t)$, while the estimated sources are obtained with $y_r(t) = B_x x(t)$.

In order to evaluate the performance of the criteria, one could think in showing their graph and their isosurfaces. However, this approach is mainly suited for evaluating the criteria over two dimensional subspaces, and it is not practical in higher dimensions. For this reason, we propose to visualize the performance of the criteria over random triangular simplicies that contain the solution. We choose two vertices of each simplex as random matrices $B_r^{(1)}, B_r^{(2)}$ that are constrained to operate on the signal subspace, while the third vertex is automatically determined by the matrix $B_r^{(3)} = 3B_r^{*} - (B_r^{(1)} + B_r^{(2)})$. The three vertex $B_r^{(1)}, B_r^{(2)}$ and $B_r^{(3)}$ are used to parameterize in barycentric coordinates the elements $B_r(\lambda_1, \lambda_2, \lambda_3) = \sum_{v=1}^{3} \lambda_v B_r^{(v)}$ of the triangular set of extraction matrices

$$B_{tri} = \left\{ B_r(\lambda_1, \lambda_2, \lambda_3) : \sum_{v=1}^{3} \lambda_v = 1, \lambda_v \in [0, 1] \right\}. \quad \text{(102)}$$

The solution $B_r^{*} = B_r(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, located at the barycenter of $B_{tri}$, recovers the target source $y_r(t) = B_x x(t) = s_r(t)$.

The considered model satisfies the BCA hypotheses H1-H6 for the target sources with $q = 2$ (circular shape), although there is a strong element-wise cross-correlation between the target and nuisance sources $\approx 0.57$.

Figure 4 presents an experimental illustration of the hierarchy of the criteria based on the correct hypothesis like $C(B_r), C_1(B_r)$ and $C_2(B_r, \{2, 2\})$, and a comparison with a criterion based on an approximate hypothesis: $C_3(B_r, \infty)$ which relays on H6 with $q = \infty$ (square shape). Figures 5(a) and 5(b) compare the graph of these criteria over the triangular set of extraction matrices $B_{tri}$. We can observe how these criteria successfully find the minimum of their respective graphs at $B_r^{*}$, which is marked with a yellow cross at the barycenter of each triangular set.

The criteria $C(B_r), C_1(B_r)$ and $C_2(B_r, \{2, 2\})$ attain the same value at their minimum because they are based on correct and coherent sets of hypotheses for the model of the observations. Also note how those criteria based on a large number of correct hypotheses are more discriminative, in the sense that their risk increases faster for a given departure from the solution. On the other hand, the criterion $C_3(B_r, \infty)$ has a greater value at its minimum (see Figure 4) because its square shape hypothesis is only a rough approximation of
Fig. 5. Figures of the first experiment: subplots a) and b) present color plots of the graphs of the criteria and their level sets over \( B_{\text{tri}} \). Symmetric positions (with respect to the vertical division between neighbor triangles) correspond to the same extraction matrix.

so that it does no longer satisfy the hypothesis H4 (this can be seen in subfigure 6(a)). The nuisance source is still defined by (101). In this second example, we can observe in subfigure 6(c) that the minimum of the criterion \( C(B_r) \) (left triangle) still finds the target source, but this is no longer the case for the criterion \( C_1(B_r) \) (right triangle) that assumes H4.

In order to get closer to assumption H4, we can try to align the target sources with the coordinate axis. This is done in our case by applying a rotation

\[
U = \frac{1}{2} \begin{pmatrix} 1 + j & 1 + j \\ 1 + j & -1 - j \end{pmatrix}
\]

(103)

to the outputs, i.e., replacing the extraction matrix \( B_r \) by \( B'_{r} = UB_r \). In spite of this alignment, which can be seen in figure 6(b), the sources still do not satisfy the hypothesis H4. However, in most cases it turns out that the resulting approximation is sufficient so as to restore the capability of \( C_1(B_r) \) to discriminate the target source, although at expense of having at the solution \( C_1(B_r) > C(B_r) \) what implies a higher minimum risk. This behavior can be observed in subfigure 6(d).

C. Separation of bounded and possibly correlated sources

The main purpose of the third experiment is to illustrate the advantages of using a BCA algorithm when the sources have clear boundaries, assumption which is especially accurate in the case of the communications signals. Additionally, the experiment illustrates how the BCA hypotheses can still hold true for correlated sources [17]-[20], justifying in this situation the preference of BCA criteria over other ones based on the strict independence of the sources.

We present an application of BCA in MIMO communications with \( n \) correlated transmitters and \( m \geq n \) receivers. The problem consist in, given only \( T \) samples of the observation vector obtained by the receiver, try to blindly estimate the
transmitted sources and the mixing matrix of the flat-fading channel. Consider $T = 500$ noisy observations $x(t) = As(t) + w(t)$, where $s(t)$ are the sources, $A$ is a matrix of i.i.d. complex Gaussian samples that models the propagation in a Rayleigh flat fading channel, and $w(t)$ denotes the realization of the additive white Gaussian noise.

We set $n = 5$ and $m = 10$ to respectively generate the sources and the observations. The standard deviation of the noise is one-tenth of the original sources, i.e. $\sigma_w = 0.1 \sigma_s$, which corresponds with an input-SNR of $20\,\text{dB}$. The five sources initially belong to a $32$-QAM constellation. Their samples are temporally independent while their spatial dependence is controlled by the parameter $\rho$. This variable represents the normalized cross-correlation between any two consecutive sources while, for sources of normalized variance ($\sigma^2_s = 1$), the correlation matrix of the vector of sources is given by the symmetric Toeplitz matrix

$$
\Sigma_s = \begin{pmatrix}
1 & \rho^1 & \rho^2 & \cdots & \rho^{n-1} \\
\rho^1 & 1 & \rho^2 & \cdots & \rho^{n-2} \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\rho^{n-1} & \cdots & \rho^1 & 1 & \rho^1
\end{pmatrix}.
$$

We implemented a bounded component analysis separation algorithm ($p = n$) based on criterion $C_1(B_\tau)$. This is a two step algorithm which projects the observations onto the signal subspace, optimizes the contrast in the Stiefel manifold and then releases the orthogonality constraint to run an unconstrained version of the same criterion.

We compare the average signal to interference plus noise ratio (SINR) performance of the bounded component analysis algorithm based on criterion $C_1(B_\tau)$ with that of two popular ICA algorithms: the Joint Approximate Diagonalization of Eigen-matrices (JADE) algorithm [40], and the non-circular implementation of FastICA (nc-FastICA) [41]. In similarity [14], we also introduce an upper-bound in the comparison to know how far the algorithms are from the theoretical (non-blind) SINR performance limit. The optimal solution in terms of SINR is given by the spatial Wiener filter trained with the $T$ samples of the sources. More details on the SINR definition and evaluation can be found in [14, pag. 2150].

For a given random mixing matrix, the average SINR of a Monte Carlo simulation over 100 independent realizations of observations is computed. The results of the simulation for 32-QAM sources are shown in subfigure 7-(a). We can observe how the BCA criterion is close to the SINR upper-bound for
most of the values of the correlation coefficient \( \rho \in [-1, 1] \), whereas, in concordance with the theory, the performance of ICA algorithms quickly deteriorates as \( \rho \) deviates from 0.

We also present in subfigure 7-(b) the experimental results of a simulation for complex uniform sources of square shape. The results are similar to those of the previous simulation, although the SINR performance is a few dB worse. This difference in performance can be explained by the fact that, in finite samples, the convergence of the sample convex support \( \mathcal{S}_y \) towards the true convex support \( \mathcal{S}_Y \) is faster for discrete sources with clear boundaries (like the 32-QAM) than for continuous uniform sources.

As a last remark, the simulations of figure 7 also evidence that using the uniform distribution of the outputs as robust strategy for the risk minimization in our admissible set, does not imply that the target sources should be close uniform in Kullback-Leibler sense. The divergence \( D(p_Y || p_{y1}) \) of the estimated sources from the uniform distribution may be high or even diverge to infinity (as in the case of \( p_{y1} \) being a discrete distribution) and still the BCA criteria can work well, as we have seen in subfigure 7-(a).

X. Conclusions

We have adopted the principle of minimum risk for solving the problem of the estimation of the bounded components of the observations in a linear mixture. This principle uses the minimization of the worst case complexity in our description of the observations as the suitable framework for connecting the evidence and the hypotheses about the sources with the criteria for their estimation. We have shown how the minimization of the complexity risk in the estimation of the density of the observations drives us towards the strategy of the minimization of a normalized volume of the desired source estimate. Depending on the geometric hypotheses for the sources, the suitable evaluation of the normalized volume varies, which can explain the existing differences between the bounded component analysis criteria. These findings have been corroborated through numerical simulations. In the future, we intend to analyze the identifiability conditions for the sources and to describe in more detail the algorithms that implement the proposed criteria.

APPENDIX A

Proof of Lemma 1

It has been already shown in [14, Lemma 1] that \( \overline{S}(x, x_\nu) = S(x) \times S(x_\nu) \) is a sufficient condition for \( \mathcal{S}_x + \mathcal{S}_x_\nu = \mathcal{S}_x \times \mathcal{S}_x_\nu \). and this result together with condition A2-b automatically implies the assumption A2. So, we only need to prove the reverse implication. Consider that A2 is true, then

\[
\mathcal{S}_x = \{ x_\tau + x_\nu : (x_\tau, x_\nu) \in \overline{S}(x, x_\nu) \}.
\]

should be equal to

\[
\mathcal{S}_x + \mathcal{S}_{x_\nu} = \{ x_\tau + x_\nu : (x_\tau, x_\nu) \in S(x) \times S(x_\nu) \}.
\]

A first necessary condition for the equality is that \( \{X_\tau\} \cap \{X_\nu\} = \{0\} \) which implies A2-b. Then, due the uniqueness of the representation, the second necessary condition for the equality is \( \mathcal{S}(x, x_\nu) = \mathcal{S}_x \times \mathcal{S}_x_\nu \), which matches with condition A2-a, thereby completing the proof.

APPENDIX B

The convexity of the admissible set

The convexity of the admissible set of distributions \( \mathcal{P}_X(\mathbf{B}, \pi_Y) \) defined in (38), is proved when, for any two densities \( p_{\nu}, p_1 \) within the set and any \( \lambda \in [0, 1] \), it is true that the mixture distribution \( p_\lambda = \lambda p_0 + (1 - \lambda) p_1 \) always
belongs to. As the convexity is preserved under linear transformations, it is sufficient to check the convexity over the set of corresponding distributions for the outputs

\[ P_Y(B, \pi_Y) = \left\{ p_Y = p_{Y_0}, p_{Y_1} | r_0 \subseteq S_{y_0}, \mu_{Y_1 | r_0} = \mu_{Y_0 | r_1}, \Sigma_{Y_1 | r_0} = \Sigma_{y_0 | r_1} \right\}. \] (107)

However, due to the independence between the random variable \( Y \) and the conditional random variable \( Y_1 | r_0 \), one only need to prove that the subsets \( A = \{ p_{Y_0} : S_{y_0} \subseteq S_{y_0} \} \) and \( B = \{ p_{Y_1} : \mu_{Y_1 | r_0} = \mu_{Y_0 | r_1}, \Sigma_{Y_1 | r_0} = \Sigma_{y_0 | r_1} \} \) are both convex.

It is easy to check that for \( p_{Y_0}, p_{Y_1} \in A \) and any \( \lambda \in [0, 1] \), the convex support of \( p_{Y_2} = \lambda p_{Y_0} + (1-\lambda) p_{Y_1} \) satisfies \( S_{y_2} = S_{y_0} \cup S_{y_1} \subseteq S_{y_0} \) and therefore \( p_{Y_2} \in A \). Similarly, for \( p_{Y_0}, p_{Y_1} \in B, i.e. \) two distributions with the same mean and covariance, for any \( \lambda \in [0, 1] \) and \( p_{Y_2} = \lambda p_{Y_0} + (1-\lambda)p_{Y_1} \), the mean and covariance are preserved due to the linearity of the expectation, thus \( p_{Y_2} \in B \), which concludes the proof.

APPENDIX C

THE MINIMUM RISK DISTRIBUTION.

We have seen in theorem 1 that the minimum risk estimate \( \hat{p}_X(x) \) is given by the distribution of maximum differential entropy within the admissible set \( P_X(B, \pi_y) \). Therefore, for the proof of lemma 3, we shall determine this distribution for the set defined in (38).

Consider the invertible correspondence between the output and the observations \( Y \in C^n \rightarrow X = B^{-1} Y \in C^n \),

and the differential entropy of the observations

\[ h(\hat{p}_X) = h(\hat{p}_Y) - \log |B B^H|. \] (109)

Since \( \hat{p}_Y = \hat{p}_{Y_0}, \hat{p}_{Y_1} | r_0 \), we can decompose the differential entropy of the output in terms of the sum of differential entropies of the factors, resulting

\[ h(\hat{p}_X) = h(\hat{p}_{Y_0}) + h(\hat{p}_{Y_1} | r_0) - \log |B B^H|. \] (110)

Those estimates that belong to the admissible set \( \hat{p}_X \in P_X(B, \pi_y) \), should be based on densities \( \hat{p}_Y \), and \( \hat{p}_{Y_1} | r_0 \) that respectively satisfy

- constraint I: \( \{ S_{y_1} \subseteq S_{y_0} \} \),
- constraint II: \( \{ \mu_{Y_1 | r_0} = \mu_{Y_0 | r_1}, \Sigma_{Y_1 | r_0} = \Sigma_{y_0 | r_1} \} \).

The uniform random variable \( Y \) is \( Y_0^N \), with support on \( S_{y_0} \), maximizes \( h(\hat{p}_{Y_0}) \) under constraint I. On the other hand, under the constraint II, the differential entropy of the conditional random variable \( Y_1 | r_0 \) is maximum for a proper conditional Gaussian random vector \( Y_1^N | Y_0^N \) with mean \( \mu_{y_1 | r_0} \) and covariance matrix \( \Sigma_{y_1 | r_0} \). Thus, the overall maximum of the differential entropy is

\[ h(\hat{p}_X) = h(p_{Y_0}) + h(p_{Y_1} | Y_0) - \log |B B^H|, \]

\[ = \int \hat{p}_X \log \frac{1}{|B B^H| p_{Y_0} p_{Y_1 | Y_0} } dY_0. \]

and the unique distribution in \( P_X(B, \pi_y) \) that can reach this maximum is given by

\[ \hat{p}_X(x) = |B B^H| p_{Y_0} p_{Y_1 | Y_0} (Y_0 | Y_0). \]

APPENDIX D

SIMPLIFYING THE MINIMUM RISK SOLUTION INTO A NEGENTROPY-LIKELIHOOD CRITERION

In order to prove the first part of the lemma 6 we shall show that, under regularity condition 1, the solution of minimum risk decomposes as

\[ R_\psi(\hat{p}_X) = \psi(p_{X^D}, p_{X^\psi}) - \Delta \psi( B_r), \] (111)

where \( \psi(p_{X^D}, p_{X^\psi}) \) is independent \( B_r \) and equal to \( \log (\pi e)^m |\Sigma_1| \), whereas \( \Delta \psi(B_r) \) represents the maximization criterion described in equation (58).

We start from the value of the minimum risk \( R_\psi(\hat{p}_X) \), which equals the complexity \( \psi(p_{X^D}, \hat{p}_X) \). Due to corollary 2 we can compute it in terms of the empirical density \( p_{X}^D \) as

\[ R_\psi(\hat{p}_X) = \psi(p_{X^D}, \hat{p}_X) = \int p_{X^D}(x) \log \frac{1}{\hat{p}_X(x)} dx. \] (112)

With the help of lemma 5 we substitute \( \hat{p}_X(x) \) to obtain

\[ R_\psi(\hat{p}_X) = \int p_{X^D}(x) \log \frac{1}{p_{Y_0}^D(x)} dx + \int p_{X^D}(x) \log \frac{p_{Y_\psi|Y_0}^D (Y_0 | Y_0)}{p_{Y_\psi|Y_0}^D (Y_0 | Y_0)} dx. \] (113)

The first term in the right-hand-side of this equation is equal to \( \psi(p_{X^D}, p_{X^\psi}) \), the complexity of the robust Gaussian approximation presented in lemma 4. Since \( p_{X^D} \) and \( p_{X^\psi} \) are both members of the family \( P_X(\{ \mu_x, \Sigma_x \}) \), the equilibrium property of this robust estimate enforces the equality

\[ \psi(p_{X^D}, p_{X^\psi}) = \psi(p_{X^\psi}, p_{X^\psi}) = \log (\pi e)^m |\Sigma_1|. \] (114)

Equating (113) and (111), we see that for the lemma to be true it is necessary that the negentropy-likelihood criterion \( \Delta \psi (B_r) \) could be written as

\[ \Delta \psi(B_r) = \int p_{Y_0}^D(x) \log \frac{p_{Y_\psi|Y_0}^D (Y_0 | Y_0)}{p_{Y_\psi|Y_0}^D (Y_0 | Y_0)} dx. \] (115)

The change of the variable of integration from \( x \) to \( y = Bx \) yields the simplification

\[ \Delta \psi(B_r) = \int p_{Y_0}^D(y) \log \frac{p_{Y_\psi|Y_0}^D (Y_0 | Y_0)}{p_{Y_\psi|Y_0}^D (Y_0 | Y_0)} dy. \] (116)

Substituting \( p_{Y_0}^D(y) = p_{Y_0}^D(y) p_{Y_\psi|Y_0}^D (Y_0 | Y_0) \) and integrating out \( p_{Y_0}^D(y) \) we obtain the desired result

\[ \Delta \psi(B_r) = \int p_{Y_0}^D(Y_0) \log \frac{p_{Y_\psi|Y_0}^D (Y_0 | Y_0)}{p_{Y_\psi|Y_0}^D (Y_0 | Y_0)} dy_0 = \psi(p_{Y_\psi}, p_{Y_\psi}) - \psi(p_{Y_\psi}, p_{Y_\psi}). \] (117)

This is the maximization criterion described in equation (58), which proves the first part of the lemma.

In order to prove the second part of the lemma, we will need to use the equilibrium in complexity for the robust Gaussian
and uniform estimates of the target output. This equilibrium enables us to write
\[
\Delta \psi(B_r) = \psi(p_{Y_r}, p_{Y_r'}) - \psi(p_{Y_r}, p_{Y_r'}). 
\]
Now, with the help of the identity
\[
D(p_{Y_r} \parallel p_{Y_r'}) = \psi(p_{Y_r}, p_{Y_r'}) - h(p_{Y_r}), 
\]
we can rewrite the maximization criterion in terms of the Kullback-Leibler divergence
\[
\Delta \psi(B_r) = \psi(p_{Y_r}, p_{Y_r'}) - h(p_{Y_r}) + h(p_{Y_r'}) - \psi(p_{Y_r}, p_{Y_r'}) = D(p_{Y_r} \parallel p_{Y_r'}) - D(p_{Y_r'} \parallel p_{Y_r'}) 
\]
and use the definition (55) to recognize it as a negentropy-likelihood criterion (59), concluding the proof of the lemma.

**APPENDIX E**

**OBTAINING THE MINIMUM NORMALIZED VOLUME CRITERION**

The main equality (64) of lemma 7 can be rewritten as
\[
C(B_r) = R(p_{Y_r}) \psi(p_{Y_r}, p_{Y_r'}) - \psi(p_{Y_r}, p_{Y_r'}) + \psi(p_{Y_r'}, p_{Y_r'}). 
\]
(120)
The proof of the lemma consists in showing how the right-hand-side of the expression simplifies to the logarithmic version of the normalized volume criterion that we presented in (63).

Substituting \( R(p_{Y_r}) \) in the previous expression with the help of equations (111) and (117) we obtain
\[
C(B_r) = \psi(p_{Y_r}, p_{Y_r'}) - \Delta \psi(B_r) = \psi(p_{Y_r}, p_{Y_r'}) + \psi(p_{Y_r'}, p_{Y_r'}) - \psi(p_{Y_r}, p_{Y_r'}), 
\]
(121)
where each of these three involved complexities is given by
\[
\Psi(p_{Y_r}, p_{Y_r'}) = \log(\pi e)^p \quad \Psi(p_{Y_r}, p_{Y_r'}) = \log(\pi e)^p + \log |\Sigma_{Y_r}|, \quad \Psi(p_{Y_r'}, p_{Y_r'}) = \log V o l C_0(y_r). 
\]
(122) (123) (124)
After substituting these expressions in (121), we obtain the logarithmic version of the normalized volume criterion
\[
C(B_r) = \log \left( \frac{Vol C_0(y_r)}{|\Sigma_{Y_r}|} \right). 
\]
(125)

**REFERENCES**


Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.


Sergio Cruces (S’93–M’99–SM’06) received the Graduate and Ph.D. degrees in telecommunications engineering from the University of Vigo, Spain, in 1994 and 1999, respectively. From 1994 to 1995 he worked as project’s engineer for the Department of Signal Theory and Communications of this university. In 1995 he joined the Department of Signal Processing and Communications, University of Seville, Spain. He has been a usual Visiting Researcher with the Laboratory of Advanced Brain Signal Processing, RIKEN Brain Science Institute, Wako City, Japan. Since 2001 he is an Associate Professor and teaches undergraduate and postgraduate courses on communication theory, digital signal processing for audio systems and information theory. His current research interests include statistical learning, information theory and their role in the analysis of the components of the observations, with emphasis on their potential applications in neuroscience, audio processing and telecommunications. Dr. Cruces has served as a Program Committee Member for several international conferences. He is an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems and forms part of the editorial board of the Journal of Computational Intelligence and Neuroscience.

Iván Durán-Díaz was born in Seville, Spain, in 1975. He received the Telecommunication Engineering degree from the University of Seville in 2001, and the Ph.D. degree from the University of Seville in 2009. Since 2002 he has been with the Radiocommunications Systems Group at the Signal Theory and Communications Dept. of the University of Seville, where he is currently an Assistant Professor. He teaches undergraduate courses on digital signal processing, radiocommunications and acoustics. His current research interests are in the area of blind source separation and latent variable analysis.