entropy

MDPI

Article

Common Information Components Analysis

Erixhen Sula * and Michael C. Gastpar

School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland; michael.gastpar@epfl.ch

* Correspondence: erixhen.sula@epfl.ch

Abstract: Wyner's common information is a measure that quantifies and assesses the commonality between two random variables. Based on this, we introduce a novel two-step procedure to construct features from data, referred to as Common Information Components Analysis (CICA). The first step can be interpreted as an extraction of Wyner's common information. The second step is a form of back-projection of the common information onto the original variables, leading to the extracted features. A free parameter γ controls the complexity of the extracted features. We establish that, in the case of Gaussian statistics, CICA precisely reduces to Canonical Correlation Analysis (CCA), where the parameter γ determines the number of CCA components that are extracted. In this sense, we establish a novel rigorous connection between information measures and CCA, and CICA is a strict generalization of the latter. It is shown that CICA has several desirable features, including a natural extension to beyond just two data sets.

Keywords: common information; dimensionality reduction; feature extraction; unsupervised; canonical correlation analysis; CCA

Citation: Sula, E.; Gastpar, M.C. Common Information Components Analysis. *Entropy* **2021**, *23*, 151. https://doi.org/10.3390/e23020151

Academic Editors: Nariman Farsad, Marco Mondelli, Morteza Mardani and Boris Ryabko

Received: 30 November 2020

Accepted: 22 January 2021

Published: 26 January 2021

1. Introduction

Understanding relations between two (or more) sets of variates is key to many tasks in data analysis and beyond. To approach this problem, it is natural to reduce each of the sets of variates separately in such a way that the reduced descriptions fully capture the *commonality* between the two sets, while suppressing aspects that are individual to each of the sets. This permits to understand the relation between the two sets without obfuscation. A popular framework to accomplish this task follows the classical viewpoint of *dimensionality reduction* and is referred to as Canonical Correlation Analysis (CCA) [1]. CCA seeks the best *linear* extraction, i.e., we consider linear projections of the original variates. Via the so-called Kernel trick, this can be extended to cover arbitrary (fixed) function classes.

Wyner's common information is a well-known and established measure of the dependence of two random variables. Intuitively, it seeks to extract a third random variable such that the two random variables are conditionally independent given the third, but at the same time that the third variable is as compact as possible. Compactness is measured in terms of the mutual information that the third random variable retains about the original two. The resulting optimization problem is not a convex problem (because the constraint set is not a convex set), and therefore, not surprisingly, closed-form solutions are rare. A natural generalization of Wyner's common information is obtained by replacing the constraint of conditional independence by a limit on the conditional mutual information. If the limit is set equal to zero, we return precisely to the case of conditional independence. Exactly like mutual information, Wyner's common information and its generalization are endowed with a clear operational meaning. They characterize the fundamental limits of data compression (in the Shannon sense) for a certain network situation.

1.1. Related Work

Connections between CCA and Wyner's common information have been explored in the past. It is well known that, for Gaussian vectors (standard, non-relaxed), Wyner's common information is attained by all of the CCA components together, see [2]. This has been further interpreted, see, e.g., [3]. Needless to say, having all of the CCA components together essentially amounts to a one-to-one transform of the original data into a new basis. It does not yet capture the idea of feature extraction or dimensionality reduction. To put our work into context, it is only the *relaxation* of Wyner's common information [4,5] that permits to conceptualize the sequential, one-by-one recovery of the CCA components, and thus, the spirit of dimensionality reduction.

CCA also appears in a number of other problems related to information measures and probabilistic models. For example, in the so-called Gaussian information bottleneck problem, the optimizing solution can be expressed in terms of the CCA components [6], and an interpretation of CCA as a (Gaussian) probabilistic model was presented in [7].

Generalizations of CCA have appeared before in the literature. The most prominent is built around maximal correlation. Here, one seeks arbitrary remappings of the original data in such a way as to maximize their correlation coefficient. This perspective culminates in the well-known *alternating conditional expectation* (ACE) algorithm [8].

Feature extraction and dimensionality reduction have a vast amount of literature attached to them, and it is beyond the scope of the present article to provide a comprehensive overview. In a part of that literature, information measures play a key role. Prominent examples are independent components analysis (ICA) [9] and the information bottleneck [10,11], amongst others. More recently, feature extraction alternations via information theory are presented in [12,13]. In [12], the estimation of Rényi's quadratic entropy is studied, whereas, in [13], standard information theoretic measures such as Kullback–Leibler divergence are used for fault diagnosis. Other slightly related feature extraction methods that perform dimensionality reduction on a single dataset include [14–20]. More concretely, in [14], a sparse Support Vector Machine (SVM) approach is used for feature extraction. In [15], feature extraction is performed via regression by using curvilinearity instead of linearity. In [16], compressed sensing is used to extract features when the data have a sparse representation. In [17], an invariant mapping method is invoked to map the high-dimensional data to low-dimensional data that is based on a neighborhood relation. In [18], feature extraction is performed for a partial learning of the geometry of the manifold. In [19], distance correlation measure (a measure with similar properties as the regular Pearson correlation coefficient) is proposed as a new feature extraction method. In [20], kernel principal component analysis is used to perform feature extraction and allow for the extraction of nonlinearities. In [21], feature extraction is done by a robust regression based approach and, in [22], a linear regression approach is used to extract features.

1.2. Contributions

The contributions of our work are the following:

- We introduce a novel suit of algorithms, referred to as CICA. These algorithms are characterized by a two-step procedure. In the first step, a relaxation of Wyner's common information is extracted. The second step can be interpreted as a form of projection of the common information back onto the original data so as to obtain the respective features. A free parameter γ is introduced to control the complexity of the extracted features.
- We establish that, for the special case where the original data are jointly Gaussian, our algorithms precisely extract the CCA components. In this case, the parameter γ determines how many of the CCA components are extracted. In this sense, we establish a new rigorous connection between information measures and CCA.
- We present initial results on how to extend CICA to more than two variates.

- Via a number of paradigmatic examples, we illustrate that, for *discrete data*, CICA gives intuitively pleasing results while other methods, including CCA, do not. This is most pronounced in a simple example with three sources described in Section 7.1.

1.3. Notation

A bold capital letter such as $\mathbf{X}$ denotes a random vector, and $\mathbf{x}$ its realization. The probability distribution function of random variable X will be denoted by p_X or $p(x)$ depending on the context. A non-bold capital letter such as K denotes a (fixed) matrix, and K^H its Hermitian transpose. Specifically, $K_{\mathbf{X}}$ denotes the covariance matrix of the random vector $\mathbf{X}$. $K_{\mathbf{XY}}$ denotes the covariance matrix between random vectors $\mathbf{X}$ and $\mathbf{Y}$. Let $\mathcal{P}$ be the set of all probability distribution, discrete or continuous depending on the context. Let us denote with I_n the identity matrix of dimension $n \times n$ and 0_n the zero matrix of dimension $n \times n$. We denote by $\mathcal{L}[f(x)]_x$ the lower convex envelope of $f(x)$ with respect to x and for random variables $\mathcal{L}[f(X)]_{p_X}$ is the lower convex envelope of $f(X)$ with respect to p_X. We denote by $h_b(x) := -x \log x - (1-x) \log 1 - x$ the binary entropy for $0 \leq x \leq 1$.

1.4. A Simple Example with Synthetic Data

To set the stage and under the guise of an informal problem statement, let us consider a simple example involving synthetic data. Specifically, we consider two-dimensional data, that is, the vectors $\mathbf{X}$ and $\mathbf{Y}$ are of length 2. The goal is to extract, separately from each of the two, a one-dimensional description in such a way as to extract the commonality between $\mathbf{X}$ and $\mathbf{Y}$ while suppressing their individual features. For simplicity, in the present artificial example, we will assume that the entries of the vectors only take value in a small finite set, namely, $\{0, 1, 2, 3\}$. To illustrate the point, we consider the following special statistical model:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} U \oplus X_2 \\ X_2 \end{pmatrix}, \tag{1}$$

and

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} U \oplus Y_2 \\ Y_2 \end{pmatrix}, \tag{2}$$

where U, X_2, and Y_2 are mutually independent uniform random variables over the set $\{0, 1, 2, 3\}$ and $\oplus$ denotes addition modulo 4.

The reason for this special statistical structure is such that it is obvious what should be extracted, namely, $\mathbf{X}$ should be reduced to U, and $\mathbf{Y}$ should also be reduced to U. This reduces both $\mathbf{X}$ and $\mathbf{Y}$ to "one-dimensional" descriptions, and these one-dimensional descriptions capture precisely the dependence between $\mathbf{X}$ and $\mathbf{Y}$. In this simple example, all the commonality between $\mathbf{X}$ and $\mathbf{Y}$ is captured by U. More formally, conditioned on U, the vectors $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent.

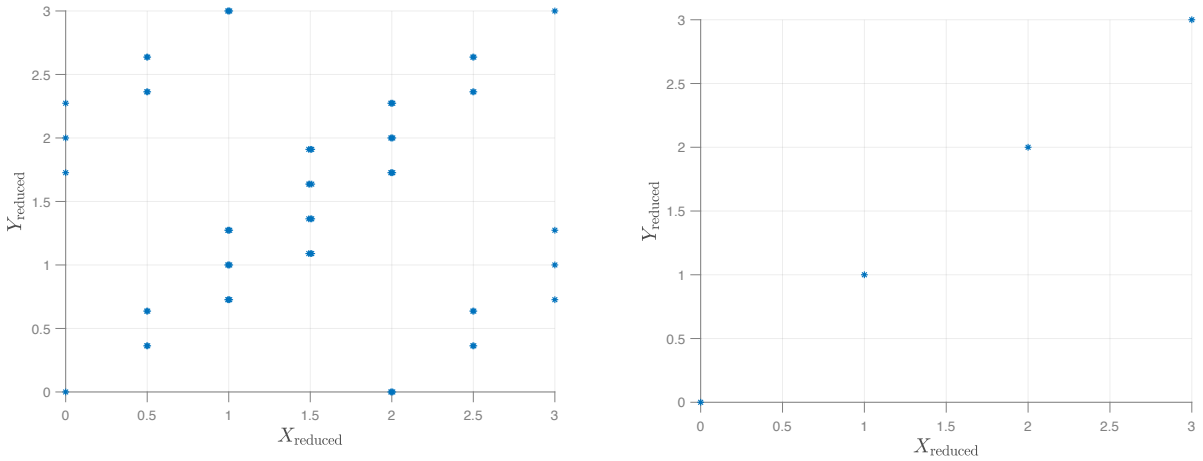
The interesting point of this example is that any *pair* of components of $\mathbf{X}$ and $\mathbf{Y}$ are *independent* of each other, such as, for example, X_1 and Y_1. Therefore, the joint covariance matrix of the merged vector $(\mathbf{X}, \mathbf{Y})$ is a scaled identity matrix. This implies that any method that only uses the covariance matrix as input, including CCA, cannot find any commonalities between $\mathbf{X}$ and $\mathbf{Y}$ in this example.

By contrast, the algorithmic procedure discussed in the present paper will correctly extract the desired answer. In Figure 1, we show numerical simulation outcomes for a couple of approaches. Specifically, in *(a)*, we can see that, in this particular example, CCA fails to extract the common features. This, of course, was done on purpose: For the synthetic data at hand, the global covariance matrix is merely a scaled identity matrix, and since CCA's only input is the covariance matrix, it does not actually do anything in this example. In *(b)*, we show the performance of the approximate gradient-descent based implementation of the CICA algorithm proposed in this paper, as detailed in Section 6. In this simple example, this precisely coincides with the ideal theoretical performance of CICA

as in a Generic Procedure 1, but, in general, the gradient-descent based implementation is not guaranteed to find the ideal solution.

At this point, we should stress that, for such a simple example, many other approaches would also lead to the same, correct answer. One of them is maximal correlation. In that perspective, one seeks to separately reduce $\mathbf{X}$ and $\mathbf{Y}$ by applying possibly nonlinear functions $f(\cdot)$ and $g(\cdot)$ in such a way as to maximize the correlation between $f(\mathbf{X})$ and $g(\mathbf{Y})$. Clearly, for the simple example at hand, selecting $f(\mathbf{X}) = X_1 \oplus X_2$ and $g(\mathbf{Y}) = Y_1 \oplus Y_2$ leads to correlation one, and is thus a maximizer.

Finally, the present example is also too simplistic to express the finer information-theoretic structure of the problem. One step up is the example presented in Section 5 below, where the commonality between $\mathbf{X}$ and $\mathbf{Y}$ is not merely an equality (the component U above), but rather a probabilistic dependency.



(**a**) CCA (**b**) CICA

Figure 1. The situation for the synthetic data as described in example described in Section 1.4. Figure 1a shows the scatterplot for two one-dimensional features extracted by CCA. Apparently, the approach is not able to extract the commonality between the vectors $\mathbf{X}$ and $\mathbf{Y}$ in this synthetic example. Figure 1b shows the performance of the heuristic algorithm of CICA described in Section 6, which, in this simple example, ends up matching the ideal theoretical performance of CICA as in a Generic Procedure 1 for $n = 10^5$ data samples.

2. Wyner's Common Information and Its Relaxation

The main framework and underpinning of the proposed algorithm is Wyner's common information and its extension, which is briefly reviewed in the sequel, along with its key properties.

2.1. Wyner's Common Information

Wyner's common information is defined for two random variables X and Y of arbitrary fixed joint distribution $p(x, y)$.

Definition 1 ([23]). *For random variables X and Y with joint distribution $p(x, y)$, Wyner's common information is defined as*

$$C(X;Y) = \inf_{p(w|x,y)} I(X,Y;W) \text{ such that } I(X;Y|W) = 0. \tag{3}$$

Wyner's common information satisfies a number of interesting properties. We state some of them below in Lemma 1 for a generalized definition given in Definition 2.

We note that explicit formulas for Wyner's common information are known only for a small number of special cases. The case of the doubly symmetric binary source is solved completely in [23] and can be written as

$$C(X;Y) = 1 + h_b(a_0) - 2h_b\left(\frac{1-\sqrt{1-2a_0}}{2}\right), \tag{4}$$

where a_0 denotes the probability that the two sources are unequal (assuming without loss of generality $a_0 \leq \frac{1}{2}$). In this case, the optimizing W in Equation (3) can be chosen to be binary. Further special cases of discrete-alphabet sources appear in [24].

Moreover, when X and Y are jointly Gaussian with correlation coefficient ρ, then $C(X;Y) = \frac{1}{2}\log\frac{1+|\rho|}{1-|\rho|}$. Note that, for this example, $I(X;Y) = \frac{1}{2}\log\frac{1}{1-\rho^2}$. This case was solved in [25,26] using a parameterization of conditionally independent distributions, and we have recently found an alternative proof that also extends to the generalization of Wyner's common information discussed in the next subsection [5].

2.2. A Natural Relaxation of Wyner's Common Information

A natural generalization of Wyner's common information (Definition 1) is obtained by replacing the constraint of conditional independence with a limit γ on the conditional mutual information, leading to the following:

Definition 2 (from [4,5,27])**.** *For random variables X and Y with joint distribution $p(x,y)$, we define*

$$C_\gamma(X;Y) = \inf_{p(w|x,y)} I(X,Y;W) \text{ such that } I(X;Y|W) \leq \gamma. \tag{5}$$

This definition appears in slightly different form in Wyner's original paper (Section 4.2 in [23]), where an auxiliary quantity $\Gamma(\delta_1,\delta_2)$ is defined satisfying $C_\gamma(X;Y) = H(X,Y) - \Gamma(0,\gamma)$. The above definition first appears in [4]. Comparing Definitions 1 and 2, we see that, for $\gamma = 0$, we have $C_0(X;Y) = C(X;Y)$, the regular Wyner's common information. In this sense, one may refer to $C_\gamma(X;Y)$ as *relaxed* Wyner's common information.

In line with the discussion following Definition 1, it is not surprising that explicit solutions to the optimization problem in Definition 2 are very rare. In fact, the only presently known general solution concerns the case of jointly Gaussian random variables [5]. The corresponding formula is given below in Theorem 1.

By contrast, the case of the doubly symmetric binary source remains open. An upper bound for this case is given by choosing the auxiliary W as

$$W = \begin{cases} X \oplus V, & \text{if } X = Y, \\ U, & \text{if } X \neq Y, \end{cases} \tag{6}$$

where V is Bernoulli with probability α and U is Bernoulli with probability $\frac{1}{2}$. Thus, the upper bound is

$$C_\gamma(X;Y) \leq I(X,Y;W) = 1 - (1-a_0)h_b(\alpha) - a_0, \tag{7}$$

where $\alpha \geq \alpha_W$ is chosen such that

$$I(X;Y|W) = 2h_b\Big((1-\alpha)(1-a_0) + \frac{a_0}{2}\Big) - (1-a_0)h_b(\alpha) - a_0 - h_b(a_0) = \gamma, \tag{8}$$

where

$$\alpha_W = \frac{(1-\sqrt{1-2a_0})^2}{4(1-a_0)}. \tag{9}$$

Numerical studies (Section 3.5 in [28]) suggest that this upper bound is tight, but no formal proof is available to date.

The following lemma summarizes some basic properties of $C_\gamma(X;Y)$.

Lemma 1 (partially from [5]). *$C_\gamma(X;Y)$ satisfies the following basic properties:*

1. $C_\gamma(X;Y) \geq \max\{I(X;Y) - \gamma, 0\}$.
2. *Data processing inequality: If $X - Y - Z$ forms a Markov chain, then $C_\gamma(X;Z) \leq \min\{C_\gamma(X;Y), C_\gamma(Y;Z)\}$.*
3. *$C_\gamma(X;Y)$ is a convex and continuous function of γ for $\gamma \geq 0$.*
4. *Tensorization: For n independent pairs $\{(X_i, Y_i)\}_{i=1}^n$, we have that*

$$C_\gamma(X^n;Y^n) = \min \sum_{i=1}^n C_{\gamma_i}(X_i;Y_i),$$

   where the min *is over all non-negative $\{\gamma_i\}_{i=1}^n$ satisfying $\sum_{i=1}^n \gamma_i = \gamma$.*
5. *If $Z - X - Y$ forms a Markov chain, then $C_\gamma((X,Z);Y) = C_\gamma(X;Y)$.*
6. *The cardinality of $\mathcal{W}$ may be restricted to $|\mathcal{W}| \leq |\mathcal{X}||\mathcal{Y}| + 1$.*
7. *If $f(\cdot)$ and $g(\cdot)$ are one-to-one functions, then $C_\gamma(f(X);g(Y)) = C_\gamma(X;Y)$.*
8. *For discrete X, we have $C_\gamma(X;X) = \max\{H(X) - \gamma, 0\}$.*

Proofs of items 1–4 are given in [5], and the proofs of items 5–8 are given in Appendix A.

2.3. The Non-Convexity of the Relaxed Wyner's Common Information Problem

It is important to observe that the optimization problem of Definition 2 is not a convex problem. First, we observe that $I(X,Y;W)$ is indeed a convex function of $p(w|x,y)$, which is a well-known fact, see, e.g., (Theorem 2.7.4 in [29]). The issue is with the constraint set. The set of distributions $p(w|x,y)$ for which $I(X;Y|W) \leq \gamma$ is not a convex set. To provide some intuition for the structure of this set, let us consider $I(X;Y|W)$ as a function of $p(w|x,y)$, and examine its (non-)convexity. The relation between the two is described by the epigraph

$$\text{epigraph}\{I(X;Y|W)\} = \{(p(w|x,y), \gamma) : p(w|x,y) \in \mathcal{P}, \gamma \geq I(X;Y|W)\}. \tag{10}$$

The function $I(X;Y|W)$ is convex in $p(w|x,y)$ if and only if its epigraph is a convex set which would imply that the set of distributions $p(w|x,y)$ for which $I(X;Y|W) \leq \gamma$ is also convex. Now, we present an example that $I(X;Y|W)$ is not a convex function of $p(w|x,y)$.

Example 1. *Let the distributions $p(x,y), p_1(w|x,y), p_2(w|x,y)$ be*

$$p(x,y) = \begin{bmatrix} \frac{2}{5} & \frac{1}{10} \\ \frac{1}{10} & \frac{2}{5} \end{bmatrix}, p_1(w|x,y) = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{3}{4} & \frac{3}{4} & \frac{3}{4} & \frac{3}{4} \end{bmatrix}, p_2(w|x,y) = \begin{bmatrix} \frac{1}{2} & \frac{3}{4} & \frac{3}{4} & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}, \tag{11}$$

respectively. For this example, one can evaluate numerically that, under $p_1(w|x,y)$, we have $I_{p_1}(X;Y|W) < 0.279$ and under $p_2(w|x,y)$, we have $I_{p_2}(X;Y|W) < 0.262$. By the same token, one can show that, under $(p_1(w|x,y) + p_2(w|x,y))/2$, we have $I_{(p_1+p_2)/2}(X;Y|W) > 0.274$. Hence, we conclude that, for this example,

$$I_{(p_1+p_2)/2}(X;Y|W) > \frac{1}{2}\big(I_{p_1}(X;Y|W) + I_{p_2}(X;Y|W)\big), \tag{12}$$

which proves that $I(X;Y|W)$ cannot be convex.

2.4. The Operational Significance of the Relaxed Wyner's Common Information Problem

It is important to note that Wyner's common information has clear and well-defined operational significance. This is perhaps not central to the detailed explanations and examples given in the sequel. However, it has a role in the appreciation of the rigorous

connection established in our work. An excellent description of the operational significance is given in (Section I in [23]), where two separate aspects are identified. The first concerns a source coding scenario with a single encoder and two decoders, one interested in X and the other in Y. Three bit streams are constructed by the encoder: One public (to both decoders), and two private streams, one for each decoder. Then, $C(X;Y)$ characterizes the minimum number of bits that must be sent on the public stream such that the total number of bits sent stays at the global minimum, which is well known to be the joint entropy of X and Y. If the rate on the public bit stream dips below $C(X;Y)$, it is no longer possible to keep the total rate at the joint entropy. Rather, there is a strict penalty now, and this penalty can be expressed via $C_\gamma(X;Y)$. The second rigorous operational significance concerns the distributed generation of correlated randomness. We have two separate processors, one generating X and the other Y. For a fixed desired resulting probability distribution $p(x,y)$, how many common random bits (shared between both processors) are required? Again, the answer is precisely $C(X;Y)$. A connection between caching and the Gray–Wyner network is developed in [30].

3. The Algorithm

The main technical result of this paper is to establish that the outcome of a specific procedure induced by the relaxed Wyner's common information is tantamount to CCA whenever the original underlying distribution is Gaussian. In preparation for this, in this section, we present the proposed algorithm. In doing so, we will assume that the distribution of the data are $p(\mathbf{x},\mathbf{y})$. In many applications involving CCA, the data distributions may not be known, but, rather, a number of samples of $\mathbf{X}$ and $\mathbf{Y}$ are provided, based on which CCA would then estimate the covariance matrix. A similar perspective can be taken on our procedure, but is left for future work. A short discussion can be found in Section 8 below.

3.1. High-Level Description

The proposed algorithm takes as input the distribution $p(\mathbf{x},\mathbf{y})$ of the data, as well as a level γ. The level γ is a non-negative real number and may be thought of as a resolution level or a measure of coarseness: If $\gamma = 0$, then the full commonality (or common information) between $\mathbf{X}$ and $\mathbf{Y}$ is extracted in the sense that it is conditioned on the common information, $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent. Conversely, if γ is large, then only the most important part of the commonality is extracted. Fixing the level γ, the idea of the proposed algorithm is to evaluate the relaxed Wyner's Common Information of Equation (5) between the information sources (data sets) at the chosen level γ. This evaluation will come with an associated conditional distribution $p_\gamma(w|x,y)$, namely, the conditional distribution attaining the minimum in the optimization problem of Equation (5). The second half of the proposed algorithm consists of leveraging the minimizing $p_\gamma(w|x,y)$ in such a way as to separately reduce $\mathbf{X}$ and $\mathbf{Y}$ to those features that best express the commonality. This may be thought of as a type of projection of the minimizing random variable W back onto $\mathbf{X}$ and $\mathbf{Y}$, respectively. For the case of Gaussian statistics, this can be made precise.

3.2. Main Steps of the Algorithm

The algorithm proposed here starts from the joint distribution of the data, $p(\mathbf{x},\mathbf{y})$. Estimates of this distribution can be obtained from data samples $\mathbf{X}^n$ and $\mathbf{Y}^n$ via standard techniques. The main steps of the procedure can then be described as follows:

Generic Procedure 1 (CICA)**.**

1. *Select a real number γ, where $0 \leq \gamma \leq I(\mathbf{X};\mathbf{Y})$. This is the compression level: A low value of γ represents low compression, and, thus, many components are retained. A high value of γ represents high compression, and, thus, only a small number of components are retained.*

2. *Solve the relaxed Wyner's common information problem,*

$$\min_{p(w|\mathbf{x},\mathbf{y})} I(\mathbf{X},\mathbf{Y};W) \text{ such that } I(\mathbf{X};\mathbf{Y}|W) \leq \gamma, \tag{13}$$

leading to an associated conditional distribution $p_\gamma(w|\mathbf{x},\mathbf{y})$.

3. *Using the conditional distribution* $p_\gamma(w|\mathbf{x},\mathbf{y})$ *found in Step 2), the dimension-reduced data sets can now be found via one of the following three variants:*

   (a) *Version 1: MAP (maximum a posteriori):*

$$u(\mathbf{x}) = \arg\max_w p_\gamma(w|\mathbf{x}), \tag{14}$$

$$v(\mathbf{y}) = \arg\max_w p_\gamma(w|\mathbf{y}). \tag{15}$$

   (b) *Version 2: Conditional Expectation:*

$$u(\mathbf{x}) = \mathbb{E}[W|\mathbf{X}=\mathbf{x}], \tag{16}$$

$$v(\mathbf{y}) = \mathbb{E}[W|\mathbf{Y}=\mathbf{y}]. \tag{17}$$

   (c) *Version 3: Marginal Integration:*

$$u(\mathbf{x}) = \int_{\mathbf{y}} p(\mathbf{y})\mathbb{E}[W|\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y}]d\mathbf{y}, \tag{18}$$

$$v(\mathbf{y}) = \int_{\mathbf{x}} p(\mathbf{x})\mathbb{E}[W|\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y}]d\mathbf{x}. \tag{19}$$

The present paper focuses on the three versions given here because, for these three versions, we can establish Theorem 2, showing that, in the case of Gaussian statistics, all three versions lead exactly to CCA. Second, we note that, for concrete examples, it is often evident which of the versions is preferable. For example, in Section 5, we consider a binary example where the associated W in Step 2 of our algorithm is also binary. In this case, Version 1 will reduce the original binary vector $\mathbf{X}$ to a binary scalar, which is perhaps the most desirable outcome. By contrast, Versions 2 and 3 require an explicit embedding of the binary example in the reals, and will reduce the original binary vector $\mathbf{X}$ to a real-valued scalar, which might not be as insightful.

4. For Gaussian, CICA Is CCA

In this section, we consider the special case where $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian random vectors. Since the mean has no bearing on either CCA or Wyner's common information, we will assume it to be zero in the sequel, without loss of generality. One key ingredient for this argument is a well-known change of basis, see, for example [2], which we will now introduce in detail. Note that the mean will not change any mutual information term, thus we assume it to be zero without a loss of generality. We first need to introduce notation for CCA. To this end, let us express the covariance matrices, as usual, in terms of their eigendecompositions as

$$K_{\mathbf{X}} = Q_x \begin{pmatrix} \Lambda_{r_X} & 0 \\ 0 & 0_{n-r_X} \end{pmatrix} Q_x^T \tag{20}$$

and

$$K_{\mathbf{Y}} = Q_y \begin{pmatrix} \Lambda_{r_Y} & 0 \\ 0 & 0_{n-r_Y} \end{pmatrix} Q_y^T, \tag{21}$$

where r_X and r_Y denote the rank of $K_{\mathbf{X}}$ and $K_{\mathbf{Y}}$, respectively. Starting from this, we define the matrices

$$K_{\mathbf{X}}^{-1/2} = Q_x \begin{pmatrix} \Lambda_{r_X}^{-1/2} & 0 \\ 0 & 0_{n-r_X} \end{pmatrix} Q_x^T \tag{22}$$

and

$$K_{\mathbf{Y}}^{-1/2} = Q_y \begin{pmatrix} \Lambda_{r_Y}^{-1/2} & 0 \\ 0 & 0_{n-r_Y} \end{pmatrix} Q_y^T, \tag{23}$$

where, for a diagonal matrix Λ with strictly positive entries, $\Lambda_{r_Y}^{-1/2}$ denotes the diagonal matrix whose diagonal entries are the reciprocals of the square roots of the entries of the matrix Λ. Using these matrices, the key step is to apply the change of basis

$$\hat{\mathbf{X}} = K_{\mathbf{X}}^{-1/2}\mathbf{X} \tag{24}$$

$$\hat{\mathbf{Y}} = K_{\mathbf{Y}}^{-1/2}\mathbf{Y}. \tag{25}$$

In the new coordinates, the covariance matrices of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$, respectively, can be shown to be

$$K_{\hat{\mathbf{X}}} = \begin{pmatrix} I_{r_X} & 0 \\ 0 & 0_{n-r_X} \end{pmatrix} \tag{26}$$

and

$$K_{\hat{\mathbf{Y}}} = \begin{pmatrix} I_{r_Y} & 0 \\ 0 & 0_{n-r_Y} \end{pmatrix}. \tag{27}$$

Moreover, we have

$$K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}} = K_{\mathbf{X}}^{-1/2} K_{\mathbf{XY}} K_{\mathbf{Y}}^{-1/2}. \tag{28}$$

Let us denote the singular value decomposition of this matrix by

$$K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}} = U\Sigma V^H. \tag{29}$$

where Σ contains, on its diagonal, the ordered singular values of this matrix, denoted by $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_n$. In addition, let us define

$$\tilde{\mathbf{X}} = U^H\hat{\mathbf{X}} \tag{30}$$

$$\tilde{\mathbf{Y}} = V^H\hat{\mathbf{Y}}, \tag{31}$$

which implies that $K_{\tilde{\mathbf{X}}} = K_{\hat{\mathbf{X}}}$, $K_{\tilde{\mathbf{Y}}} = K_{\hat{\mathbf{Y}}}$, and $K_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}} = \Sigma$.

Next, we will leverage this change of basis to establish Wyner's common information and its relaxation for the Gaussian vector case, and then to prove the connection between Generic Procedure 1 and CCA.

4.1. Wyner's Common Information and Its Relaxation in the Gaussian Case

For the case where $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian random vectors, a full and explicit solution to the optimization problem of Equation (5) is found in [5]. To give some high-level intuition, the proof starts by mapping from $\mathbf{X}$ to $\tilde{\mathbf{X}}$ and from $\mathbf{Y}$ to $\tilde{\mathbf{Y}}$, as in Equations (30) and (31). This preserves all mutual information expressions as well as joint Gaussianity. Moreover, due to the structure of the covariance matrices of the vectors $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$, we have that $\{(\tilde{X}_i, \tilde{Y}_i)\}_{i=1}^n$ are n independent pairs of Gaussian random variables. Thus, by the tensorization property (see Lemma 1), the vector problem can be reduced to n parallel scalar problems. The solution of the scalar problem is the main technical contribution of [5],

and we refer to that paper for the detailed proof. The resulting formula can be expressed as in the following theorem.

Theorem 1 (from [5]). *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be jointly Gaussian random vectors of length n and covariance matrix* $K_{(\mathbf{X},\mathbf{Y})}$. *Then,*

$$C_\gamma(\mathbf{X};\mathbf{Y}) = \min_{\gamma_i:\sum_{i=1}^n \gamma_i = \gamma} \sum_{i=1}^n C_{\gamma_i}(X_i;Y_i), \tag{32}$$

where

$$C_{\gamma_i}(X_i;Y_i) = \frac{1}{2}\log^+ \frac{(1+\rho_i)(1-\sqrt{1-e^{-2\gamma_i}})}{(1-\rho_i)(1+\sqrt{1-e^{-2\gamma_i}})} \tag{33}$$

and ρ_i *(for* $i = 1,\ldots,n$*) are the singular values of* $K_{\mathbf{X}}^{-1/2}K_{\mathbf{XY}}K_{\mathbf{Y}}^{-1/2}$, *where* $K_{\mathbf{X}}^{-1/2}$ *and* $K_{\mathbf{Y}}^{-1/2}$ *are defined to mean that only the positive eigenvalues are inverted.*

As pointed out above, we refer to (Theorem 7 in [5]) for a rigorous proof of this theorem.

4.2. CICA in the Gaussian Case and the Exact Connection with CCA

In this section, we consider the proposed CICA algorithm in the special case where the data distribution is $p(\mathbf{x},\mathbf{y})$, a (multivariate) Gaussian distribution. We establish that, in this case, the classic CCA is a solution to all versions of the proposed CICA algorithm. In this sense, CICA is a strict generalization of CCA. CCA is briefly reviewed in Appendix B. Leveraging the matrices U and V defined via the singular value decomposition in Equation (29), CCA performs the dimensonality reduction

$$u(\mathbf{x}) = U_k^H\hat{\mathbf{x}} = U_k^H K_{\mathbf{X}}^{-1/2}\mathbf{x} \tag{34}$$

$$v(\mathbf{y}) = V_k^H\hat{\mathbf{y}} = V_k^H K_{\mathbf{Y}}^{-1/2}\mathbf{y}, \tag{35}$$

where the matrix U_k contains the first k columns of U (that is, the k left singular vectors corresponding to the largest singular values), and the matrix V_k the respective right singular vectors. We refer to these as the "top k CCA components."

Theorem 2. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be jointly Gaussian random vectors. Then:*

1. *The top k CCA components are a solution to* **all three** *versions of Generic Procedure 1.*
2. *The parameter* γ *controls the number k as follows:*

$$k(\gamma) = \begin{cases} n, & \text{if } 0 \le \gamma < ng(\rho_n), \\ n-1, & \text{if } ng(\rho_n) \le \gamma < (n-1)g(\rho_{n-1}) + g(\rho_n), \\ n-2, & \text{if } (n-1)g(\rho_{n-1}) + g(\rho_n) \le \gamma \\ & \quad < (n-2)g(\rho_{n-2}) + g(\rho_{n-1}) + g(\rho_n), \\ \vdots, & \vdots, \\ \ell & \text{if } (\ell+1)g(\rho_{\ell+1}) + \sum_{i=\ell+2}^n g(\rho_i) \le \gamma \\ & \quad < \ell g(\rho_\ell) + \sum_{i=\ell+1}^n g(\rho_i), \\ \vdots, & \vdots, \\ 1, & \text{if } 2g(\rho_2) + \sum_{i=2}^n g(\rho_i) \le \gamma < \sum_{i=1}^n g(\rho_i), \\ 0, & \text{if } \sum_{i=1}^n g(\rho_i) \le \gamma, \end{cases} \tag{36}$$

where $g(\rho) = \frac{1}{2}\log\frac{1}{1-\rho^2}$.

Remark 1. *Note that $k(\gamma)$ is a decreasing, integer-valued function. An illustration for a special case is given in Figure 2.*

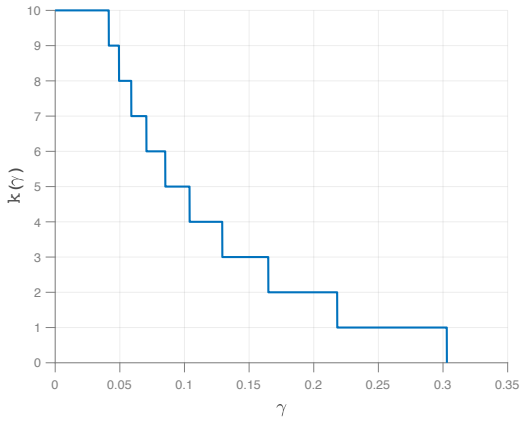


Figure 2. Illustration of the function $k(\gamma)$ from Theorem 2 for the concrete case where $\mathbf{X}$ and $\mathbf{Y}$ have $n = 10$ components each and the correlation coefficients are $\rho_m = 1/(m+1)$.

Proof. The main contribution of the theorem is the first item, *i.e.*, the connection between CCA and Generic Procedure 1 in the case where $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian. The proof follows along the steps of the CICA procedure: We first show that, in Step 2, when $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian, then the minimizing W may be taken jointly Gaussian with $\mathbf{X}$ and $\mathbf{Y}$. Then, we establish that, in Step 3, with the W from Step 2, we indeed obtain that the dimension-reduced representations $u(\mathbf{x})$ and $v(\mathbf{y})$ turn into the top k CCA components. In detail:

Step 2 of Generic Procedure 1: The technical heavy lifting for this step in the case where $p(\mathbf{x}, \mathbf{y})$ is a multivariate Gaussian distribution is presented in [5]. We shall briefly summarize it here. In the case of Gaussian vectors, the solution to the optimization problem in Equation (5) is most easily described in two steps. First, we apply the change of basis indicated in Equations (24) and (25). This is a one-to-one transform, leaving all information expressions in Equation (5) unchanged. In the new basis, we have n independent pairs. By the tensorization property (see Lemma 1), when $\mathbf{X}$ and $\mathbf{Y}$ consist of independent pairs, the solution to the optimization problem in Equation (5) can be reduced to n separate scalar optimizations. The remaining crux then is solving the scalar Gaussian version of the optimization problem in Equation (5). This is done in (Theorem 3 in [5]) via an argument of factorization of convex envelope. The full solution to the optimization problem is given in Equations (32) and (33). The remaining allocation problem over the non-negative numbers γ_i can be shown to lead to a water-filling solution, given in (Theorem 8 in [5]). More explicitly, to understand this solution, start by setting $\gamma = I(\mathbf{X}; \mathbf{Y})$. Then, the corresponding $C_\gamma(\mathbf{X}; \mathbf{Y}) = 0$ and the optimizing distribution $p_\gamma(w|\mathbf{x}, \mathbf{y})$ trivializes. Now, as we lower γ, the various terms in the sum in Equation (32) start to become non-zero, starting with the term with the largest correlation coefficient ρ_1. Hence, an optimizing distribution $p_\gamma(w|\mathbf{x}, \mathbf{y})$ can be expressed as $\mathbf{W}_\gamma = U_k^H K_\mathbf{X}^{-1/2}\mathbf{X} + V_k^H K_\mathbf{Y}^{-1/2}\mathbf{Y} + \mathbf{Z}$, where the matrices U_k and V_k are precisely the top k CCA components (see Equations (34) and (35) and the following discussion), and $\mathbf{Z}$ is additive Gaussian noise with mean zero, independent of $\mathbf{X}$ and $\mathbf{Y}$.

Step 3 of Generic Procedure 1: For the algorithm, we need the corresponding conditional

marginals, $p_\gamma(w|\mathbf{x})$ and $p_\gamma(w|\mathbf{y})$. By symmetry, it suffices to prove one formula. Changing basis as in Equations (24) and (25), we can write

$$\begin{aligned}
\mathbb{E}[W|\mathbf{X}] &= \mathbb{E}[U_k^H\hat{\mathbf{X}} + V_k^H\hat{\mathbf{Y}} + \mathbf{Z}|\hat{\mathbf{X}}] && (37)\\
&= U_k^H\hat{\mathbf{X}} + V_k^H\mathbb{E}[\hat{\mathbf{Y}}|\hat{\mathbf{X}}] && (38)\\
&= U_k^H\hat{\mathbf{X}} + V_k^H\left(\mathbb{E}[\hat{\mathbf{Y}}\hat{\mathbf{X}}^H]\left(\mathbb{E}[\hat{\mathbf{X}}\hat{\mathbf{X}}^H]\right)^{-1}\hat{\mathbf{X}}\right) && (39)\\
&= U_k^H\hat{\mathbf{X}} + V_k^H K_{\hat{\mathbf{Y}}\hat{\mathbf{X}}}\hat{\mathbf{X}} && (40)\\
&= U_k^H\hat{\mathbf{X}} + \left(K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}V_k\right)^H\hat{\mathbf{X}}. && (41)
\end{aligned}$$

The first summand contains exactly the top k CCA components extracted from $\mathbf{X}$, which is the claimed result. The second summand requires further scrutiny. To proceed, we observe that, for CCA, the projection vectors obey the relationship (see Equation (A12))

$$\mathbf{u} = \alpha K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}\mathbf{v}, \tag{42}$$

for some real-valued constant α. Thus, combining the top k CCA components, we can write

$$U_k = DK_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}V_k, \tag{43}$$

where D is a diagonal matrix. Hence,

$$\begin{aligned}
\mathbb{E}[W|\mathbf{X}] &= U_k^H\hat{\mathbf{X}} + D^{-1}U_k^H\hat{\mathbf{X}} && (44)\\
&= \tilde{D}U_k^H\hat{\mathbf{X}}, && (45)
\end{aligned}$$

where $\tilde{D}$ is the diagonal matrix

$$\tilde{D} = I + D^{-1}. \tag{46}$$

This is precisely the top k CCA components (note that the solution to the CCA problem (A7) is only specified up to a scaling). This establishes the theorem for the case of Version 2) of the proposed algorithm. Clearly, it also establishes that $p_\gamma(w|\mathbf{x})$ is a Gaussian distribution with mean given by (45), thus establishing the theorem for Version 1) of the proposed algorithm. The proof for Version 3 follows along similar lines and is thus omitted. □

5. A Binary Example

In this section, we carry through a theoretical study of a somewhat more general case of the example discussed in Section 1.4 that is believed to be within the reach of practical data. In order to do a theoretical study, we need to constrain the data into binary for the reason that computing the Wyner's common information for doubly binary symmetric source is already known.

Let us illustrate the proposed algorithm via a simple example. Consider the vector (U, X_2, V, Y_2) of binary random variables. Suppose that (U, V) is a doubly symmetric binary source. This means that U is uniform and V is the result of passing U through a binary symmetric ("bit-flipping") channel with flip probability denoted by a_0 to match the notation in (Section 3 in [23]). Without loss of generality, we may assume $a_0 \leq \frac{1}{2}$. Meanwhile, X_2 and Y_2 are independent binary uniform random variables, also independent of the pair (U, V). We will then form the vectors $\mathbf{X}$ and $\mathbf{Y}$ as

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} U \oplus X_2 \\ X_2 \end{pmatrix}, \tag{47}$$

and

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} V \oplus Y_2 \\ Y_2 \end{pmatrix}, \tag{48}$$

where $\oplus$ denotes the modulo-reduced addition, as usual. How do various techniques perform for this example?

- Let us first analyze the behavior and outcome of CCA in this particular example. The key observation is that any pair, amongst the four entries in these two vectors, X_1, X_2, Y_1, and Y_2, are (pairwise) independent binary uniform random variables. Hence, the overall covariance matrix of the merged random vector $(\mathbf{X}^T, \mathbf{Y}^T)^T$ is merely a scaled identity matrix. This, in turn, implies that CCA as described in Equations (34) and (35) merely boils down to the identity mapping. Concretely, this means that, for CCA, in this example, the best one-dimensional projections are ex aequo any pair of one coordinate of the vector $\mathbf{X}$ with one coordinate of the vector $\mathbf{Y}$. As we have already explain above, any such pair is merely a pair of independent (and identically distributed) random variables, so CCA does not extract any dependence between $\mathbf{X}$ and $\mathbf{Y}$ at all. Of course, this is the main point of the present example.
- How does CICA perform in this example? We selected this example because it represents one of the only cases for which a closed-form solution to the optimization problem in Equation (13) is known, at least in the case $\gamma = 0$. To see this, let us first observe that, in our example, we have

$$p(u, v, x_2, y_2) = p(u, v)p(x_2)p(y_2). \tag{49}$$

Next, we observe that

$$\begin{aligned} C_\gamma(\mathbf{X}; \mathbf{Y}) &= C_\gamma(U, X_2; V, Y_2) && (50)\\ &= C_\gamma(U; V, Y_2) && (51)\\ &= C_\gamma(U; V) && (52) \end{aligned}$$

where (51) follows from Lemma 1, Item 5, and the Markov chain $X_2 - U - (V, Y_2)$ that is satisfied from (49). The last equation (52) follows from Lemma 1, Item 5, and the Markov chain $Y_2 - V - U$ that is satisfied from (49). That is, in this simple example, solving the optimization problem of Equation (13) is tantamount to solving the optimization problem in Equation (52). For the latter, the solution is well known, see (Section 3 in [23]). Specifically, we can express the conditional distribution $p_\gamma(w|\mathbf{x}, \mathbf{y})$ that solves the optimization problem of Equation (13) and is required for Step 3 of Generic Procedure 1 as follows:

$$p_{\gamma=0}(w|\mathbf{x}, \mathbf{y}) = \begin{cases} 1-\nu, & \text{if } w = 0, x_1 \oplus x_2 = 0, y_1 \oplus y_2 = 0, \\ \nu, & \text{if } w = 1, x_1 \oplus x_2 = 0, y_1 \oplus y_2 = 0, \\ \nu, & \text{if } w = 0, x_1 \oplus x_2 = 1, y_1 \oplus y_2 = 1, \\ 1-\nu, & \text{if } w = 1, x_1 \oplus x_2 = 1, y_1 \oplus y_2 = 1, \\ \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{53}$$

where

$$\nu = \frac{1}{2} - \frac{\sqrt{1 - 2a_0}}{2(1 - a_0)}. \tag{54}$$

Let us now apply Version 1 (the MAP version) of Generic Procedure 1. To this end, we

also need to calculate $p_\gamma(w|\mathbf{x})$ and $p_\gamma(w|\mathbf{y})$. Again, for $\gamma = 0$, these can be expressed in a closed form as follows:

$$p_{\gamma=0}(w|\mathbf{x}) = \begin{cases} 1-a_1, & \text{if } w=0, x_1 \oplus x_2 = 0, \\ a_1, & \text{if } w=1, x_1 \oplus x_2 = 0, \\ a_1, & \text{if } w=0, x_1 \oplus x_2 = 1, \\ 1-a_1, & \text{if } w=1, x_1 \oplus x_2 = 1, \end{cases} \tag{55}$$

where

$$a_1 = \frac{1}{2}\left(1 - \sqrt{1-2a_0}\right). \tag{56}$$

The formula for $p_\gamma(w|\mathbf{y})$ follows by symmetry and shall be omitted. The final step is to follow Equations (14) and (15) and find $\arg\max_w p_{\gamma=0}(w|\mathbf{x})$ for each $\mathbf{x}$ as well as $\arg\max_w p_{\gamma=0}(w|\mathbf{y})$ for each $\mathbf{y}$. For the example at hand, these can be compactly expressed as

$$u(\mathbf{x}) = \arg\max_w p_\gamma(w|\mathbf{x}) = x_1 \oplus x_2 = u, \tag{57}$$

$$v(\mathbf{y}) = \arg\max_w p_\gamma(w|\mathbf{y}) = y_1 \oplus y_2 = v, \tag{58}$$

from the fact that $0 \le a_0 \le \frac{1}{2}$ that implies $0 \le a_1 \le \frac{1}{2}$. Hence, we find that, for CICA as described in Generic Procedure 1, an optimal solution is to reduce $\mathbf{X}$ to U and $\mathbf{Y}$ to V. This captures all the dependence between the vectors $\mathbf{X}$ and $\mathbf{Y}$, which appears to be the most desirable outcome.

As a final note, we point out that it is conceptually straightforward to evaluate Versions 2 and 3 (conditional expectation) of Generic Procedure 1 in this example, but this would require embedding the considered binary alphabets into the real numbers. This makes it a less satisfying option for the simple example at hand.

6. A Gradient Descent Based Implementation

As we discussed above, in our problem, the objective is indeed a convex function of the optimization variables (but the constraint set is not convex). Clearly, this gives hope that gradient-based techniques may lead to interesting solutions. In this section, we examine a first tentative implementation and check it against ground truth for some simple examples.

In theory for convex problems, gradient descent will guarantee convergence to the optimal solution; otherwise, it will guarantee only local convergence. Gradient descent runs in iterative steps, where each step does a local linear approximation and the step size depends on a learning parameter that is α for our problem. In our work, we want to minimize the objective $I(W;X,Y)$ when the constraint $I(X;Y|W)$ is held below a $\gamma-$level.

Instead, we apply a variant of gradient descent where we minimize the weighted sum of objective $I(W;X,Y)$ and the constraint $I(X;Y|W)$, which is $I(W;X,Y) + \lambda I(X;Y|W)$. The parameter λ will permit some control on the constraint, thus sweeping all its possible values. We present the algorithm where $C(p(w|x,y))$ will be a function of $p(w|x,y)$ that will represent $I(W;X,Y)$, and $J(p(w|x,y))$ will be a function of $p(w|x,y)$ that will represent $I(X;Y|W)$.

The exact computation of the stated update step is presented in the following lemma.

Lemma 2 (Computation of the update step)**.** *Let $p(x,y)$ be a fixed distribution, then the updating steps for the gradient descent are*

$$\frac{\partial C(p(w|x,y)}{\partial p(w|x,y)} = p(x,y)\log\frac{p(w|x,y)}{\sum_{x',y'} p(x',y')p(w|x',y')}, \tag{59}$$

$$\frac{\partial J(p(w|x,y)}{\partial p(w|x,y)} = p(x,y)\log\frac{p(w|x,y)\sum_{x',y'} p(x',y')p(w|x',y')}{\sum_{x''} p(w|x'',y)p(x''|y)\sum_{y''} p(w|x,y'')p(y''|x)}. \tag{60}$$

Proof. Let the function C be as defined above

$$C(p(w|x,y)) = \sum_{x,y,w} p(w|x,y)p(x,y)\log\frac{p(w|xy)}{\sum_{x',y'} p(w|x',y')p(x',y')}, \tag{61}$$

and, in terms of information theoretic terms, the function is $C(p(w|x,y)) = I(W;X,Y)$. In addition, $C(p(w|x,y))$ is a convex function of $p(w|x,y)$, shown in (Theorem 2.7.4 in [29]). Taking the first derivative, we get

$$\begin{aligned}\frac{\partial C(p(w|x,y))}{\partial p(w|x,y)} &= p(x,y)\log\frac{p(w|x,y)}{\sum_{x',y'} p(w|x',y')p(x',y')} + p(w|x,y)p(x,y)\frac{1}{p(w|x,y)} \\ &\quad - \sum_{x'',y''} p(w|x'',y'')p(x'',y'')\frac{p(x,y)}{\sum_{x',y'} p(w|x',y')p(x',y')} \qquad (62)\\ &= p(x,y)\log\frac{p(w|x,y)}{\sum_{x',y'} p(w|x',y')p(x',y')}. \qquad (63)\end{aligned}$$

On the other hand, the term $I(X;Y|W)$ can be expressed as

$$\begin{aligned}I(X;Y|W) &= I(W;X,Y) - I(W;X) - I(W;Y) + I(X;Y) \qquad (64)\\ &= C(p(w|x,y)) - C(p(w|x)) - C(p(w|y)) + I(X;Y). \qquad (65)\end{aligned}$$

Taking the derivative with respect to $p(w|x,y)$ becomes easier once $I(X;Y|W)$ is written in terms of function C and we already know the derivative of C from (63). Thus, the derivative would be

$$\begin{aligned}\frac{\partial J(p(w|x,y)}{\partial p(w|x,y)} &= \frac{\partial C(p(w|x,y))}{\partial p(w|x,y)} - \frac{\partial C(p(w|x))}{\partial p(w|x,y)} - \frac{\partial C(p(w|y))}{\partial p(w|x,y)} \qquad (66)\\ &= \frac{\partial C(p(w|x,y))}{\partial p(w|x,y)} - \frac{\partial C(p(w|x))}{\partial p(w|x)}\frac{\partial p(w|x)}{\partial p(w|x,y)} - \frac{\partial C(p(w|y))}{\partial p(w|y)}\frac{\partial p(w|y)}{\partial p(w|x,y)} \qquad (67)\\ &= p(x,y)\log\frac{p(w|x,y)}{\sum_{x',y'} p(w|x',y')p(x',y')} - p(x)\log\frac{p(w|x)}{\sum_{x''} p(w|x'')p(x'')}p(y|x) \\ &\quad - p(y)\log\frac{p(w|y)}{\sum_{y''} p(w|y'')p(y'')}p(x|y) \qquad (68)\\ &= p(x,y)\log\frac{p(w|x,y)\sum_{x',y'} p(x',y')p(w|x',y')}{\sum_{x''} p(w|x'',y)p(x''|y)\sum_{y''} p(w|x,y'')p(y''|x)}. \qquad (69)\end{aligned}$$

where (67) is an application of the chain rule, and the rest is straightforward computation. □

Remark 2. *In practice, it is useful and computationally cheaper to replace the derivative formulas in Lemma 2 by their standard approximations. That is, the updating step in line 7 of Algorithm 1 is replaced by*

$$\begin{aligned}\frac{\partial C(p(w|x,y))}{\partial p(w|x,y)} &\approx \frac{C(p(w|x,y)+\Delta) - C(p(w|x,y))}{\Delta}, \qquad (70)\\ \frac{\partial J(p(w|x,y))}{\partial p(w|x,y)} &\approx \frac{J(p(w|x,y)+\Delta) - J(p(w|x,y))}{\Delta}, \qquad (71)\end{aligned}$$

for a judicious choice of Δ. *This is the version that was used for Figure 1b, with* $\Delta = 10^{-3}$. *We point out that, in the general case, the error introduced by this approximation is not bounded.*

Algorithm 1: Approximate CICA Algorithm via Gradient Descent

Set $\alpha, \lambda, error$;
$\beta = \lambda \cdot \alpha$;
Initialise $p(w|x,y)$ *randomly* ;
Initialise $C_{new} \leftarrow 1, C_{old} \leftarrow 0$;
while $|C_{new} - C_{old}| > error$ **do**
  $C_{old} \leftarrow C_{new}$;
  $p(w|x,y) \leftarrow p(w|x,y) + \alpha \frac{\partial C(p(w|x,y))}{\partial p(w|x,y)} + \beta \frac{\partial J(p(w|x,y))}{\partial p(w|x,y)}$; `// update step`
  $C_{new} \leftarrow C(p(w|x,y))$;
Output $C_\gamma \leftarrow C_{new}, \gamma \leftarrow J(p(w|x,y))$;
Function $C(p(w|x,y)) \leftarrow \sum_{x,y,w} p(w|x,y)p(x,y) \log \frac{p(w|x,y)}{\sum_{x',y'} p(x',y')p(w|x',y')}$;
// $I(W;X,Y)$
Function
$J(p(w|x,y)) \leftarrow \sum_{x,y,w} p(w|x,y)p(x,y) \log \frac{p(w|x,y)p(x,y)\sum_{x',y'} p(x',y')p(w|x',y')}{\sum_{x''} p(w|x'',y)p(x'',y)\sum_{y''} p(w|x,y'')p(y'',x)}$;
// $I(X;Y|W)$

7. Extension to More than Two Sources

It is unclear how one would extend CCA to more than two databases. By contrast, for CICA, this extension is conceptually straightforward. For Wyner's common information, in Definition 1, it suffices to replace the objective in the minimization by $I(X_1, X_2, \ldots, X_M; W)$ and to keep the constraint of conditional independence. To obtain an interesting algorithm, we now need to relax the constraint of conditional independence. The most natural way is via the conditional version of Watanabe's total correlation [31], leading to the following definition:

Definition 3 (Relaxed Wyner's Common Information for M variables)**.** *For a fixed probability distribution $p(x_1, x_2, \ldots, x_M)$, we define*

$$C_\gamma(X_1; X_2; \ldots; X_M) = \inf I(X_1, X_2, \ldots, X_M; W) \tag{72}$$

such that $\sum_{i=1}^M H(X_i|W) - H(X_1, X_2, \ldots, X_M|W) \leq \gamma$, where the infimum is over all probability distributions $p(w, x_1, x_2, \ldots, x_M)$ with marginal $p(x_1, x_2, \ldots, x_M)$.

Not surprisingly, an explicit closed-form solution is difficult to find. One simple case appears below as part of the example presented in Section 7.1, see Lemma 4. By analogy with Lemma 1, we can again state basic properties.

Lemma 3. *$C_\gamma(X_1; X_2; \ldots; X_M)$ satisfies the following basic properties:*

1. $C_\gamma(X_1; X_2; \ldots; X_M) \geq \frac{1}{M-1} \max\{\sum_{i=1}^M H(X_i) - H(X_1, X_2, \ldots, X_M) - \gamma, 0\}$.
2. *$C_\gamma(X_1; X_2; \ldots; X_M)$ is a convex and continuous function of γ for $\gamma \geq 0$.*
3. *If $Z - X_1 - (X_2, \ldots, X_M)$ forms a Markov chain,*
   then $C_\gamma((X_1, Z); X_2; \ldots; X_M) = C_\gamma(X_1; X_2; \ldots; X_M)$.
4. *The cardinality of $\mathcal{W}$ may be restricted to $|\mathcal{W}| \leq \prod_{i=1}^M |\mathcal{X}_i| + 1$.*
5. *If $f_i(\cdot)$ are one-to-one functions,*
   then $C_\gamma(f_1(X_1); f_2(X_2); \ldots; f_M(X_M)) = C_\gamma(X_1; X_2; \ldots; X_M)$.
6. *For discrete X, we have $C_\gamma(X; X; \ldots; X) = \max\{H(X) - \frac{\gamma}{M-1}, 0\}$.*

Proofs for these basic properties can be found in Appendix C.

Leveraging Definition 3, it is conceptually straightforward to extend CICA (that is, Generic Procedure 1) to the case of M databases as follows. For completeness, we include an explicit statement of the resulting procedure.

Generic Procedure 2 (CICA with multiple sources)**.**

1. *Select a real number* γ*, where* $0 \leq \gamma \leq \sum_{i=1}^{M} H(\mathbf{X}_i) - H(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M)$*. This is the compression level: A low value of* γ *represents low compression, and, thus, many components are retained. A high value of* γ *represents high compression, and, thus, only a small number of components are retained.*
2. *Solving the relaxed Wyner's common information problem,*

$$\min_{p(w|\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_M)} I(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M; W) \text{ such that } \sum_{i=1}^{M} H(\mathbf{X}_i|W) - H(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M|W) \leq \gamma, \tag{73}$$

   leading to an associated conditional distribution $p_\gamma(w|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M)$*.*
3. *Using the conditional distribution* $p_\gamma(w|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M)$ *found in Step 2, the dimension-reduced data sets can now be found via one of the following three variants:*

   (a) *Version 1: MAP (maximum a posteriori):*

$$u_i(\mathbf{x}_i) = \arg\max_w p_\gamma(w|\mathbf{x}_i), \tag{74}$$

   for $i = 1, 2, \ldots, M$*.*

   (b) *Version 2: Conditional Expectation:*

$$u_i(\mathbf{x}_i) = \mathbb{E}[W|\mathbf{X}_i = \mathbf{x}_i] \tag{75}$$

   for $i = 1, 2, \ldots, M$*.*

   (c) *Version 3: Marginal Integration:*

$$u_i(\mathbf{x}_i) = \int_{\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},\mathbf{x}_{i+1},\ldots,\mathbf{x}_M} p(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_M) \mathbb{E}[W|\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_M = \mathbf{x}_M] d\mathbf{x}_1 \cdots d\mathbf{x}_{i-1} d\mathbf{x}_{i+1} \cdots d\mathbf{x}_M \tag{76}$$

   for $i = 1, 2, \ldots, M$*.*

Clearly, Generic Procedure 2 closely mirrors Generic Procedure 1. The key difference is that there is no direct analog of Theorem 2. This is no surprise since it is unclear how CCA would be extended to beyond the case of two sources. Nonetheless, it would be very interesting to explore what Generic Procedure 2 boils down to in the special case when all vectors are jointly Gaussian. At the current time, this is unknown. In fact, the explicit solution to the optimization problem in Definition 3 is presently an open problem.

Instead, we illustrate the promise of Generic Procedure 2 via a simple binary example in the next section. The example mirrors some of the basic properties of the example tackled in Section 5.

7.1. A Binary Example with Three Sources

In this section, we develop an example with three sources that borrows some of the ideas from the example discussed in Section 5. In a sense, the present example is even more illustrative because, in it, *any* two of the original vectors $\mathbf{X}_1, \mathbf{X}_2$, and $\mathbf{X}_3$, are (pairwise) independent. Therefore, any method based on pairwise measures, including CCA and maximal correlation, would not identify any commonality at all. Specifically, we consider the following simple statistical model:

$$\mathbf{X}_1 = \begin{pmatrix} U \\ Z_1 \end{pmatrix}, \mathbf{X}_2 = \begin{pmatrix} V \\ Z_2 \end{pmatrix}, \mathbf{X}_3 = \begin{pmatrix} U \oplus V \\ Z_3 \end{pmatrix}, \tag{77}$$

where U, V, Z_1, Z_2, Z_3 are independent uniform binary variables and $\oplus$ denotes modulo-2 addition. We observe that, amongst these three vectors, any pair is independent. This

implies, for example, that any correlation-based technique (including maximal correlation) will not identify any relevant features, since correlation is a pairwise measure. By contrast, we can show that one output of Algorithm 2 is indeed to select $W = (U, V)$, for $\gamma = 0$. Thus, the algorithm would reduce each of the three vectors to their first component, which is the intuitively pleasing answer in this case. By going through the steps of the Generic Procedure 2, for $\gamma = 0$, where the the joint distribution satisfies

$$p(u, v, u \oplus v, z_1, z_2, z_3) = p(u, v, u \oplus v)p(z_1)p(z_2)p(z_3) \tag{78}$$

we have that

$$\begin{aligned} C(\mathbf{X}_1; \mathbf{X}_2; \mathbf{X}_3) &= C(U, Z_1; V, Z_2; U \oplus V, Z_3) && (79)\\ &= C(U; V, Z_2; U \oplus V, Z_3) && (80)\\ &= C(U; V; U \oplus V, Z_3) && (81)\\ &= C(U; V; U \oplus V) && (82) \end{aligned}$$

where we use Lemma 3, Item 3, together with the Markov chain $Z_1 - U - (Z_2, V, Z_3, U \oplus V)$ that follows from (78) to prove step (80). Similarly, the Markov chain $Z_2 - V - (U, Z_3, U \oplus V)$ proves step (81) by making use of Lemma 3, Item 3. A similar argument is used for the last step (82). Managing to compute $C(U; V; U \oplus V)$ is equivalent to computing $C(\mathbf{X}_1; \mathbf{X}_2; \mathbf{X}_3)$, and we demonstrate how to compute it in the next part.

Lemma 4. *Let U, V be independent uniform binary variables and $\oplus$ denotes modulo-2 addition. Then, the optimal solution to*

$$C_{\gamma=0}(U; V; U \oplus V) = \inf_{W: H(U|W)+H(V|W)+H(U\oplus V|W)-H(U,V,U\oplus V|W)=0} I(W; U, V, U \oplus V) \tag{83}$$

is $W = (U, V)$, where the expression evaluates to two.

The proof is given in Appendix D. If we apply Version 1 of Step 3 of Generic Procedure 2, we obtain

$$\arg\max_w p_{\gamma=0}(w|\mathbf{x}_1) = \{(u, 0), (u, 1)\}, \tag{84}$$

that is, in this case, the maximizer is not unique. However, as we observe that the set of maximizers is a deterministic function of u alone, it is natural to reduce as follows:

$$u_1(\mathbf{x}_1) = u. \tag{85}$$

By the same token, we can reduce

$$\begin{aligned} u_2(\mathbf{x}_2) &= v, && (86)\\ u_3(\mathbf{x}_3) &= u \oplus v. && (87) \end{aligned}$$

In this example, it is clear that this indeed extracts all of the dependency there is between our three sources, and, thus, is the correct answer.

As pointed out above, in this simple example, any pair of the random vectors $\mathbf{X}_1, \mathbf{X}_2$, and $\mathbf{X}_3$ are (pairwise) independent, which implies that the classic tools based on pairwise measures (CCA, maximal correlation) cannot identify any commonality between $\mathbf{X}_1, \mathbf{X}_2$, and $\mathbf{X}_3$.

8. Conclusions and Future Work

We introduce a novel two-step procedure that we refer to as CICA. The first step consists of an information minimization problem related to Wyner's common information,

while the second can be thought of as a type of back-projection. We prove that, in the special case of Gaussian statistics, this two-step procedure precisely extracts the CCA components. A free parameter γ in CICA permits selecting the number of CCA components that are being extracted. In this sense, the paper establishes a novel rigorous connection between CCA and information measures. A number of simple examples are presented. It is also shown how to extend the novel algorithm to more than two sources.

Future work includes a more in-depth study and consideration to assess the practical promise of this novel algorithm. This will also require moving beyond the current setting where it was assumed that the probability distribution of the data at hand was provided directly. Instead, this distribution has to be estimated from data, and one needs to understand what limitations this additional constraint will end up imposing.

Author Contributions: Conceptualization, M.C.G.; Data curation, M.C.G.; Formal analysis, E.S. and M.C.G.; Methodology, M.C.G.; Software, E.S.; Validation, E.S.; Visualization, E.S. and M.C.G.; Writing—original draft, E.S. and M.C.G.; Writing—review & editing, E.S. and M.C.G. The authors have contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This work was supported in part by the Swiss National Science Foundation under Grant No. 169294.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available from the corresponding author on request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CCA	Canonical correlation analysis
ACE	Alternating conditional expectation
ICA	Independent component analysis
CICA	Common information component analysis

Appendix A. Proof of Lemma 1

For Item 5, on the one hand, we have

$$C_\gamma((X,Z);Y) = \inf_{p(w|x,y,z):I(X,Z;Y|W)\leq\gamma} I(X,Z,Y;W) \tag{A1}$$

$$\leq \inf_{p(w|x,y):I(X;Y|W)+I(Z;Y|W,X)\leq\gamma} I(X,Y;W) + I(Z;W|X,Y) \tag{A2}$$

$$= C_\gamma(X;Y) \tag{A3}$$

where, in Equation (A2), we add the constraint that conditioned on (X,Y), W is selected to be *independent* of Z, which cannot reduce the value of the infimum. That is, for such a choice of W, we have the Markov chain $Z-(X,Y)-W$, thus $I(Z;W|X,Y)=0$. Furthermore, observe that the factorization $p(x,y,z,w)=p(x,y)p(z|x)p(w|x,y)$ also implies the factorization $p(x,y,z,w)=p(x,w)p(z|x)p(y|w,x)$. Hence, we also have the Markov chain

$Z-(W,X)-Y$; thus, $I(Z;Y|W,X)=0$, which thus established the last step. Conversely, observe that

$$C_\gamma((X,Z);Y) = \inf_{p(w|x,y,z):I(X,Z;Y|W)\leq\gamma} I(X,Y,Z;W) \tag{A4}$$

$$\geq \inf_{p(w|x,y):I(X;Y|W)\leq\gamma} I(X,Y;W) + \inf_{p(w|x,y,z):I(X,Z;Y|W)\leq\gamma} I(Z;W|X,Y) \tag{A5}$$

$$\geq C_\gamma(X;Y) \tag{A6}$$

where (A5) follows from the fact that the infimum of the sum is lower bounded by the sum of the infimums and the fact that relaxing constraints cannot increase the value of the infimum, and (A6) follows from non-negativity of the second term.

Item 6 is a standard cardinality bound, following from the arguments in [32]. For the context at hand, see also Theorem 1 in (p. 6396, [33]). Item 7 follows because all involved mutual information terms are invariant to one-to-one transforms. For Item 8), note that we can express $C_\gamma(X;X)=H(X)-\max_{p(w|x):H(X|W)\leq\gamma} H(X|W)$, which directly gives the result.

Appendix B. A Brief Review of Canonical Correlation Analysis (CCA)

A brief review of CCA [1] is presented. Let $\mathbf{X}$ and $\mathbf{Y}$ be zero-mean real-valued random vectors with covariance matrices $K_\mathbf{X}$ and $K_\mathbf{Y}$, respectively. Moreover, let $K_{\mathbf{XY}}=\mathbb{E}[\mathbf{X}\mathbf{Y}^H]$. We first apply the change of basis as in (24) and (25). CCA seeks to find vectors $\mathbf{u}$ and $\mathbf{v}$ such as to maximize the correlation between $\mathbf{u}^H\hat{\mathbf{X}}$ and $\mathbf{v}^H\hat{\mathbf{Y}}$, that is,

$$\max_{\mathbf{u},\mathbf{v}} \frac{\mathbb{E}[\mathbf{u}^H\hat{\mathbf{X}}\hat{\mathbf{Y}}^H\mathbf{v}]}{\sqrt{\mathbb{E}[|\mathbf{u}^H\hat{\mathbf{X}}|^2]}\sqrt{\mathbb{E}[|\mathbf{v}^H\hat{\mathbf{Y}}|^2]}}, \tag{A7}$$

which can be rewritten as

$$\max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^H K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}\mathbf{v}}{\|\mathbf{u}\|\ \|\mathbf{v}\|}, \tag{A8}$$

where

$$K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}} = K_\mathbf{X}^{-1/2} K_{\mathbf{XY}} K_\mathbf{Y}^{-1/2}. \tag{A9}$$

Note that this expression is *invariant* to arbitrary (separate) scaling of $\mathbf{u}$ and $\mathbf{v}$. To obtain a unique solution, we could choose to impose that both vectors be unit vectors,

$$\max_{\mathbf{u},\mathbf{v}:\|\mathbf{u}\|=\|\mathbf{v}\|=1} \mathbf{u}^H K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}\mathbf{v}. \tag{A10}$$

From Cauchy–Schwarz, for a fixed $\mathbf{u}$, the maximizing (unit-norm) $\mathbf{v}$ is given by

$$\mathbf{v} = \frac{K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}^H\mathbf{u}}{\left\|K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}^H\mathbf{u}\right\|}, \tag{A11}$$

or, equivalently, for a fixed $\mathbf{v}$, the maximizing (unit-norm) $\mathbf{u}$ is given by

$$\mathbf{u} = \frac{K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}\mathbf{v}}{\left\|K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}\mathbf{v}\right\|}. \tag{A12}$$

Plugging in the latter, we obtain

$$\max_{\mathbf{v}:\|\mathbf{v}\|=1} \frac{\mathbf{v}^H K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}^H K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}\mathbf{v}}{\left\|K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}\mathbf{v}\right\|}, \tag{A13}$$

or, dividing through,

$$\max_{\mathbf{v}:\|\mathbf{v}\|=1} \left\| K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}}\mathbf{v} \right\|. \tag{A14}$$

The solution to this problem is well known: $\mathbf{v}$ is the right singular vector corresponding to the largest singular vector of the matrix $K_{\hat{\mathbf{X}}\hat{\mathbf{Y}}} = K_{\mathbf{X}}^{-1/2} K_{\mathbf{XY}} K_{\mathbf{Y}}^{-1/2}$. Evidently, $\mathbf{u}$ is the corresponding left singular vector. Restarting again from Equation (A7), but restricting to vectors that are orthogonal to the optimal choices of the first round leads to the second CCA components, and so on.

Appendix C. Proof of Lemma 3

For item 1, we proceed as follows

$$C_\gamma(X_1; X_2; \ldots; X_M) = \inf_{W: H(X_1|W)+H(X_2|W)+\cdots+H(X_M|W)-H(X_1,X_2,\ldots,X_M|W)\leq\gamma} I(W; X_1, X_2, \ldots, X_M) \tag{A15}$$

$$\geq \inf_W L(\lambda, p(w|x_1, x_2, \ldots, x_M)) \tag{A16}$$

where we used weak duality for $\lambda \geq 0$ and $L(\lambda, p(w|x_1, x_2, \ldots, x_M)$ is

$$\begin{aligned} L(\lambda, p(w|x_1, x_2, \ldots, x_M) := I(W; X_1, X_2, \ldots, X_M) + \lambda[H(X_1|W) \\ + H(X_2|W) + \cdots + H(X_M|W) - H(X_1, X_2, \ldots, X_M|W) - \gamma]. \end{aligned} \tag{A17}$$

By setting $\lambda = \frac{1}{M-1}$, we obtain

$$L(\frac{1}{M-1}, p(w|x_1, x_2, \ldots, x_M)) \tag{A18}$$

$$= \frac{M}{M-1} I(W; X_1, X_2, \ldots, X_M) - \frac{1}{M-1}[I(W; X_1) + I(W; X_2) + \cdots + I(W; X_M)]$$
$$+ \frac{1}{M-1}[H(X_1) + H(X_2) + \cdots + H(X_M) - H(X_1, X_2, \ldots, X_M) - \gamma] \tag{A19}$$

$$= \frac{1}{M-1}[I(W; X_2, \ldots, X_M|X_1) + \cdots + \frac{1}{M-1}[I(W; X_1, \ldots, X_{M-1}|X_M)$$
$$+ \frac{1}{M-1}[H(X_1) + H(X_2) + \cdots + H(X_M) - H(X_1, X_2, \ldots, X_M) - \gamma], \tag{A20}$$

where the infimum of $L(\frac{1}{M-1}, p(w|x_1, x_2, \ldots, x_M))$ in (A20) is attained for the trivial random variable W, thus $C_\gamma(X_1; X_2; \ldots; X_M) \geq \frac{1}{M-1}[H(X_1) + H(X_2) + \cdots + H(X_M) - H(X_1, X_2, \ldots, X_M) - \gamma]$. Item 2 follows from a similar argument as in (Corollary 4.5 in [23]). For item 3, we start by showing both sides of the inequality that will result in equality. One side of the inequality is shown below:

$$C_\gamma(X_1, Z; X_2; \ldots; X_M) \tag{A21}$$

$$= \inf_{W: H(X_1,Z|W)+H(X_2|W)+\cdots+H(X_M|W)-H(X_1,Z,X_2,\ldots,X_M|W)\leq\gamma} I(W; X_1, Z, X_2, \ldots, X_M) \tag{A22}$$

$$= \inf_{\substack{W: H(X_1|W)+H(X_2|W)+\cdots+H(X_M|W)-H(X_1,X_2,\ldots,X_M|W) \\ +I(Z;X_2,\ldots,X_M|X_1,W)\leq\gamma}} I(W; X_1, X_2, \ldots, X_M) + I(W; Z|X_1, X_2, \ldots, X_M) \tag{A23}$$

$$\leq C_\gamma(X_1; X_2; \ldots; X_M) \tag{A24}$$

where the last inequality follows by restricting the possible set of W, such that W and Z are conditionally independent given $(X_1, X_2, \ldots, X_M)$,

$$I(Z; W|X_1, X_2, \ldots, X_M) = 0. \tag{A25}$$

From the statement of the lemma, we have $Z - X_1 - (X_2, \ldots, X_M)$,

$$I(Z; X_2, \ldots, X_M|X_1) = 0. \tag{A26}$$

By adding (A25) and (A26), we get $I(Z; W, X_2, \ldots, X_M|X_1) = 0$. This implies that we have $I(Z; X_2, \ldots, X_M|X_1, W) = 0$, which appears in the constraint of (A23). For the other part of the inequality we proceed as follows:

$$C_\gamma(X_1, Z; X_2; \ldots; X_M) \tag{A27}$$

$$= \inf_{\substack{W: H(X_1|W)+H(X_2|W)+\cdots+H(X_M|W)-H(X_1,X_2,\ldots,X_M|W) \\ +I(Z;X_2,\ldots,X_M|X_1,W)\leq\gamma}} I(W; X_1, X_2, \ldots, X_M) + I(W; Z|X_1, X_2, \ldots, X_M) \tag{A28}$$

$$\geq C_\gamma(X_1; X_2; \ldots; X_M), \tag{A29}$$

where the last part follows by relaxing the constraint set as $I(Z; X_2, \ldots, X_M|X_1, W) \geq 0$ and, by further bounding the terms in the objective, $I(W; Z|X_1, X_2, \ldots, X_M) \geq 0$.

Item 4 is a standard cardinality bound, following from a similar argument in [32]. Item 5 follows because all involved mutual information terms are invariant to one-to-one transforms. For item 6, we apply the definition of relaxed Wyner's common information for M variables, and we have

$$C_\gamma(X; X; \ldots; X) = \inf_{W:(M-1)H(X|W)\leq\gamma} I(X; W) \tag{A30}$$

$$= H(X) - \sup_{W:(M-1)H(X|W)\leq\gamma} H(X|W) \tag{A31}$$

$$= H(X) - \frac{\gamma}{M-1}. \tag{A32}$$

Appendix D. Proof of Lemma 4

An upper bound to the problem is to pick $W = (U, V)$, thus

$$C(U; V; U \oplus V) \leq H(U, V, U \oplus V) = 2. \tag{A33}$$

Another equivalent way of writing the problem is by splitting the constraint into two constraints, as we already know that the constraint cannot be smaller than zero, so it has to be exactly zero, and it can be written in the following way:

$$C(U; V; U \oplus V) = \inf_{\substack{W: I(U\oplus V; U, V|W)=0 \\ I(U;V|W)=0}} I(W; U, V, U \oplus V). \tag{A34}$$

By using weak duality for $\lambda \geq 0$, a lower bound to the problem would be the following

$$C(U; V; U \oplus V) \geq \inf_{W: U-W-V} I(W; U, V, U \oplus V) + \lambda[H(U, V|W) + H(U \oplus V|W) - H(U, V, U \oplus V|W)]. \tag{A35}$$

By further using the constraint $U - W - V$, the above expression can be written as

$$C(U;V;U\oplus V) \geq \inf_{W:U-W-V} I(W;U,V,U\oplus V) + \lambda[H(U|W)+H(V|W)+H(U\oplus V|W) - H(U,V,U\oplus V|W)] \tag{A36}$$

$$= H(U,V,U\oplus V) + \inf_{W:U-W-V} \lambda[H(U|W)+H(V|W)+H(U\oplus V|W)] - (1+\lambda)H(U,V,U\oplus V|W) \tag{A37}$$

$$\geq H(U,V,U\oplus V) + \inf_{\tilde{U},\tilde{V}} \inf_{\substack{W:\tilde{U}-W-\tilde{V} \\ \tilde{U}\oplus\tilde{V}-(\tilde{U},\tilde{V})-W}} \lambda[H(\tilde{U}|W)+H(\tilde{V}|W)+H(\tilde{U}\oplus\tilde{V}|W)] - (1+\lambda)H(\tilde{U},\tilde{V},\tilde{U}\oplus\tilde{V}|W) \tag{A38}$$

$$= 2 + \inf_{\tilde{U},\tilde{V}} \underbrace{\inf_{\substack{W:\tilde{U}-W-\tilde{V} \\ \tilde{U}\oplus\tilde{V}-(\tilde{U},\tilde{V})-W}} \lambda H(\tilde{U}\oplus\tilde{V}|W) - H(\tilde{U}|W) - H(\tilde{V}|W)}_{\mathcal{L}[\lambda H(\tilde{U}\oplus\tilde{V})-H(\tilde{U})-H(\tilde{V})]_{p_{\tilde{U}}p_{\tilde{V}}}} \tag{A39}$$

where (A38) is a consequence of allowing a minimization (if minimum exists) over binary random variables $\tilde{U},\tilde{V}$ and the rest of equalities is straightforward manipulation. The last equation is in terms of the lower convex envelope with respect to the distribution $p_{\tilde{U}}p_{\tilde{V}}$. The aim is to search for the tightest bound over λ by studying the lower convex envelope with respect to $p_{\tilde{U}}p_{\tilde{V}}$, which, for binary $\tilde{U},\tilde{V}$, can be simplified into

$$\mathcal{L}[\lambda H(\tilde{U}\oplus\tilde{V}) - H(\tilde{U}) - H(\tilde{V})]_{p_{\tilde{U}}p_{\tilde{V}}} = \mathcal{L}[\lambda h_b(\alpha\beta+(1-\alpha)(1-\beta)) - h_b(\alpha) - h_b(\beta)]_{\alpha,\beta} \tag{A40}$$

and the latter function is a lower convex envelope with respect to $0\leq\alpha,\beta\leq 1$. Note that (A40) is a continuous function of α,β, so a first order and a second order differentiation will be enough to compute the lower convex envelope. As a result for $\lambda\geq 2$, the lower convex envelope of the right-hand side of (A40) is just zero, thus completing the proof.

References

1. Hotelling, H. Relations between two sets of variants. *Biometrika* **1936**, *28*, 321–377. [CrossRef]
2. Satpathy, S.; Cuff, P. Gaussian secure source coding and Wyner's common information. In Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, 14–19 June 2015; pp. 116–120.
3. Huang, S.L.; Wornell, G.W.; Zheng, L. Gaussian universal features, canonical correlations, and common information. In Proceedings of the 2018 IEEE Information Theory Workshop (ITW), Guangzhou, China, 25–29 November 2018.
4. Gastpar, M.; Sula, E. Relaxed Wyner's common information. In Proceedings of the 2019 IEEE Information Theory Workshop, Visby, Sweden, 25–28 August 2019.
5. Sula, E.; Gastpar, M. On Wyner's common information in the Gaussian Case. *arXiv* **2019**, arXiv:1912.07083.
6. Chechik, G.; Globerson, A.; Tishby, N.; Weiss, Y. Information bottleneck for Gaussian variables. *J. Mach. Learn. Res.* **2005**, *6*, 165–188.
7. Bach, F.; Jordan, M. *A Probabilistic Interpretation of Canonical Correlation Analysis*; Technical Report 688; University of California: Berkeley, CA, USA, 2005.
8. Breiman, L.; Friedman, J.H. Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.* **1985**, *80*, 580–598. [CrossRef]
9. Comon, P. Independent component analysis. In Proceedings of the International Signal Processing Workshop on High-Order Statistics, Chamrousse, France, 10–12 July 1991; pp. 111–120.
10. Witsenhausen, H.S.; Wyner, A.D. A conditional entropy bound for a pair of discrete random variables. *IEEE Trans. Inf. Theory* **1975**, *21*, 493–501. [CrossRef]
11. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 22–24 September 1999; pp. 368–377.
12. Xiao-Tong Yuan, B.H.; Hu, B.G. Robust feature extraction via information theoretic learning. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 1193–1200.

13. Wang, H.; Chen, P. A feature extraction method based on information theory for fault diagnosis of reciprocating machinery. *Sensors* **2009**, *9*, 2415–2436. [CrossRef]
14. Bi, J.; Bennett, K.P.; Embrechts, M.; Breneman, C.M.; Song, M. Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res.* **2003**, *3*, 1229–1243.
15. Laparra, V.; Malo, J.; Camps-Valls, G. Dimensionality reduction via regression in hyperspectral imagery. *IEEE J. Sel. Top. Signal Process.* **2015**, *9*, 1026–1036. [CrossRef]
16. Gao, J.; Shi, Q.; Caetano. Dimensionality reduction via compressive sensing. *Pattern Recognit. Lett.* **2012**, *33*, 1163–1170. [CrossRef]
17. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006.
18. Zhang, Z.; Zha, H. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Sci. Comput.* **2004**, *26*, 313–338. [CrossRef]
19. Vepakomma, P. Supervised dimensionality reduction via distance correlation maximization. *Electron. J. Stat.* **2018**, *12*, 960–984. [CrossRef]
20. Wu, J.; Wang, J.; Liu, L. Feature extraction via KPCA for classification of gait patterns. *Hum. Mov. Sci.* **2007**, *26*, 393–411. [CrossRef] [PubMed]
21. Lai, Z.; Mo, D.; Wong, W.K.; Xu, Y.; Miao, D.; Zhang, D. Robust discriminant regression for feature extraction. *IEEE Trans. Cybern.* **2018**, *48*, 2472–2484.
22. Wang, H.; Zhang, Y.; Waytowich, N.R.; Krusienski, D.J.; Zhou, G.; Jin, J.; Wang, X.; Cichocki, A. Discriminative feature extraction via multivariate linear regression for SSVEP-based BCI. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2016**, *24*, 532–541. [CrossRef] [PubMed]
23. Wyner, A. The common information of two dependent random variables. *IEEE Trans. Inf. Theory* **1975**, *21*, 163–179. [CrossRef]
24. Witsenhausen, H.S. Values and bounds for the common information of two discrete random variables. *SIAM J. Appl. Math* **1976**, *31*, 313–333. [CrossRef]
25. Xu, G.; Liu, W.; Chen, B. Wyner's common information for continuous random variables—A lossy source coding interpretation. In Proceedings of the Annual Conference on Information Sciences and Systems, Baltimore, MD, USA, 23–25 March 2011.
26. Xu, G.; Liu, W.; Chen, B. A lossy source coding interpretation of Wyner's common information. *IEEE Trans. Inf. Theory* **2016**, *62*, 754–768. [CrossRef]
27. Gastpar, M.; Sula, E. Common information components analysis. In Proceedings of the Information Theory and Applications Workshop (ITA), San Diego, CA, USA, 2–7 February 2020.
28. Wang, C.Y. Function Computation over Networks: Efficient Information Processing for Cache and Sensor Applications. Ph.D. Thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2015.
29. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2005.
30. Timo, R.; Bidokhti, S.S.; Wigger, M.A.; Geiger, B.C. A rate-distortion approach to caching. *IEEE Trans. Inf. Theory* **2018**, *64*, 1957–1976. [CrossRef]
31. Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.* **1960**, *4*, 66–82. [CrossRef]
32. Ahlswede, R.; Körner, J. Source coding with side information and a converse for degraded broadcast channels. *IEEE Trans. Inf. Theory* **1975**, *21*, 629–637. [CrossRef]
33. Wang, C.Y.; Lim, S.H.; Gastpar, M. Information-theoretic caching: Sequential coding for computing. *IEEE Trans. Inf. Theory* **2016**, *62*, 6393–6406. [CrossRef]