Journal of Machine Learning Research 18 (2018) 1-29 Submitted 7/17; Revised 11/17; Published 4/18

Gaussian Lower Bound for the Information Bottleneck Limit

Amichai Painsky AMICHAI.PAINSKY@MAIL.HUJI.AC.IL

Naftali Tishby TISHBY@CS.HUJI.AC.IL

School of Computer Science and Engineering and
The Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem
Givat Ram, Jerusalem 91904, Israel

Editor: Samuel Kaski

Abstract

The Information Bottleneck (IB) is a conceptual method for extracting the most compact, yet informative, representation of a set of variables, with respect to the target. It generalizes the notion of minimal sufficient statistics from classical parametric statistics to a broader information-theoretic sense. The IB curve defines the optimal trade-off between representation complexity and its predictive power. Specifically, it is achieved by minimizing the level of mutual information (MI) between the representation and the original variables, subject to a minimal level of MI between the representation and the target. This problem is shown to be in general NP hard. One important exception is the multivariate Gaussian case, for which the Gaussian IB (GIB) is known to obtain an analytical closed form solution, similar to Canonical Correlation Analysis (CCA). In this work we introduce a Gaussian lower bound to the IB curve; we find an embedding of the data which maximizes its "Gaussian part", on which we apply the GIB. This embedding provides an efficient (and practical) representation of any arbitrary data-set (in the IB sense), which in addition holds the favorable properties of a Gaussian distribution. Importantly, we show that the optimal Gaussian embedding is bounded from above by non-linear CCA. This allows a fundamental limit for our ability to Gaussianize arbitrary data-sets and solve complex problems by linear methods.

Keywords: Information Bottleneck, Canonical Correlations, ACE, Gaussianization, Mutual Information Maximization, Infomax

1. Introduction

The problem of extracting the relevant aspects of complex data is a long standing staple in statistics and machine learning. The Information Bottleneck (IB) method, presented by Tishby et al. (1999), approaches this problem by extending its classical notion to a broader information-theoretic setup. Specifically, given the joint distribution of a set of explanatory variables $\underline{X}$ and a target variable $\underline{Y}$ (which may also be of a higher dimension), the IB method strives to find the most compressed representation of $\underline{X}$, while preserving information about $\underline{Y}$. Thus, $\underline{Y}$ implicitly regulates the compression of $\underline{X}$, so that its compressed representation maintains a level of relevance as explanatory variables with respect to $\underline{Y}$.

The IB problem is formally defined as follows:

$$\begin{aligned} \min_{P(\underline{T}|\underline{X})} \quad & I(\underline{X};\underline{T}) \\ \text{subject to} \quad & I(\underline{T};\underline{Y}) \geq I_Y \end{aligned} \tag{1}$$

where $\underline{T}$ is the compressed representation of $\underline{X}$ and the minimization is over the mapping of $\underline{X}$ to $\underline{T}$, defined by the conditional probability $P(\underline{T}|\underline{X})$. Here, I_Y is a constant parameter that sets the level of information to be preserved between the compressed representation and the target. Solving this problem for a range of I_Y values defines the *IB curve* – a continuous concave curve which provides the optimal trade-off between representation complexity (regarded as $I(\underline{X};\underline{T})$) and predictive power ($I(\underline{T};\underline{Y})$).

The IB method showed to be a powerful tool in a variety of machine learning domains and related areas (Slonim and Tishby, 2000; Friedman et al., 2001; Sinkkonen and Kaski, 2002; Slonim et al., 2005; Hecht et al., 2009). It is also applicable to other fields such as neuroscience (Schneidman et al., 2001) and optimal control (Tishby and Polani, 2011). Recently, Tishby and Zaslavsky (2015) and Shwartz-Ziv and Tishby (2017) demonstrated its abilities in analyzing and optimizing the performance of deep neural networks.

Generally speaking, solving the IB problem (1) for an arbitrary joint distribution is not a simple task. In the introduction of the IB method, Tishby et al. (1999) defined a set of self-consistent equations which formulate the necessary conditions for the optimal solution of (1). Further, they provided an iterative Arimoto– Blahut like algorithm which shows to converge to local optimum. In general, these equations do not hold a tractable solution and are usually approximated by different means (Slonim, 2002). An extensive attention was given to the simpler categorical setup, where the IB curve is somewhat easier to approximate. Here, $\underline{X}$ and $\underline{Y}$ take values on a finite set and $\underline{T}$ represents (soft and informative) clusters of $\underline{X}$ (Slonim, 2002). Naturally, the IB problem also applies for continuous variables. In this case, approximating the solution to the self-consistent equations is even more involved. A special exception is the Gaussian case, where $\underline{X}$ and $\underline{Y}$ are assumed to follow a jointly normal distribution and the *Gaussian IB* problem (GIB) is analytically solved by linear projections to the canonical correlation vector space (Chechik et al., 2005). However, evaluating the IB curve for arbitrary continuous random variables is still considered a highly complicated task where most attempts focus on approximating or bounding it (Rey and Roth, 2012; Chalk et al., 2016). A detailed discussion regarding currently known methods is provided in the following section.

In this work we present a novel Gaussian lower bound to the IB curve, which applies to all types of random variables (continuous, nominal and categorical). Our bound strives to maximize the "jointly Gaussian part" of the data and apply the analytical GIB to it. Specifically, we seek for two transformations, $\underline{U} = \phi(\underline{X})$ and $\underline{V} = \psi(\underline{Y})$ so that $\underline{U}$ and $\underline{V}$ are highly correlated and "as jointly Gaussian as possible". In addition, we ask that the transformations preserve as much information as possible between $\underline{X}$ and $\underline{Y}$. This way, we maximize the portion of the data that can be explained by linear means, $I(\underline{U};\underline{V}) \leq I(\underline{X};\underline{Y})$, specifically using the GIB.

In fact, our results go beyond the specific context of the information bottleneck. In this work we tackle the fundamental question of linearizing non-linear problems. In other words, we ask ourselves whether it is possible to "push" all the information in the data

to its second moments. This problem has received a great amount of attention over the years. For example, Schneidman et al. (2006) discuss this problem in the context of neural networks; they provide preliminary evidence that in the vertebrate retina, weak pairwise correlations may describe the collective (non-linear) behavior of neurons. In this work, we provide both fundamental limits and constructive algorithms for maximizing the part of the data that can be optimally analyzed by linear means. This basic property holds both theoretical and practical implications, as it defines the maximal portion which allows favorable analytical properties in many applications. Interestingly, we show that even if we allow the transformations $\underline{U} = \phi(\underline{X})$ and $\underline{V} = \psi(\underline{Y})$ to increase the dimensions of $\underline{X}$ and $\underline{Y}$, our ability to linearize the problem is still limited, and governed by the non-linear canonical correlations (Breiman and Friedman, 1985) of the original variables.

Our suggested approach may also be viewed as an extension of the *Shannon lower bound* (Cover and Thomas, 2012), for evaluating the mutual information. In his seminal work, Shannon provided an analytical Gaussian lower bound for the generally involved rate distortion function. He showed that the rate distortion function $R(D)$ can be bounded from below by $h(X) - \frac{1}{2}\log(2\pi eD)$ where X is the compressed source, $h(X)$ is its corresponding deferential entropy and $\frac{1}{2}\log(2\pi eD)$ is the differential entropy of an independent Gaussian noise with a maximal distortion level D. This bound holds some favorable theoretical properties (Cover and Thomas, 2012) and serves as one of the most basic tools for approximating the rate distortion function to this very day. In this work, we use a similar rationale and derive a Gaussian lower bound for the mutual information of two random variables, which holds an analytical expression just like the Shannon's bound. We then extend our result to the entire IB curve and discuss its theoretical properties and practical considerations. A matlab implementation of our suggested approach is publicly available at the first author's web-page[1].

The rest of this manuscript is organized as follows: In Section 2 we review previous work on the IB method for continuous random variables. Section 3 defines our suggested lower bound and formulates it as an optimization problem. We then propose a set of solutions and bounds to this problem, as we distinguish between the easier univariate case (Section 4) and the more involved multivariate case (Section 5). Finally, in Section 6 we extend our results to the entire IB curve.

2. Related work

As discussed in the previous section, solving the IB problem for continuous variables is in general a difficult task. A special exception is where $\underline{X}$ and $\underline{Y}$ follow a jointly normal distribution. Chechik et al. (2005) show that in this case, the Gaussian IB problem (GIB) is solved by a noisy linear projection, $T = A\underline{X} + \underline{\zeta}$. Specifically, assume that $\underline{X}$ and $\underline{Y}$ are of dimensions n_X and n_Y respectively and denote the covariance matrix of $\underline{X}$ as $C_{\underline{X}}$ while the conditional covariance matrix of $\underline{X}|\underline{Y}$ is $C_{\underline{X}|\underline{Y}}$. Then, $\underline{\zeta}$ is a Gaussian random vector with a zero mean and a unit covariance matrix, independent of $\underline{X}$. The matrix A is defined

1. `https://sites.google.com/site/amichaipainsky/software`

as follows:

$$A = \left\{ \begin{array}{cc} [0^T; \ldots; 0^T] & 0 \leq \beta \leq \beta_1^C \\ [a_1 v_1^T; 0^T; \ldots; 0^T] & \beta_1^C \leq \beta \leq \beta_2^C \\ [a_1 v_1^T; a_2 v_2^T; 0^T; \ldots; 0^T] & \beta_2^C \leq \beta \leq \beta_3^C \\ \vdots & \vdots \end{array} \right\}. \tag{2}$$

where $\{v_1^T, v_2^T, \ldots, v_{n_x}^T\}$ are the left eigenvectors of $C_{\underline{X}|\underline{Y}} C_{\underline{X}}^{-1}$, sorted by their corresponding ascending eigenvalues $\lambda_1, \ldots, \lambda_{n_X}$, $\beta_I^C = \frac{1}{\lambda_i}$ are the critical β values, a_i are defined by $a_i = \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$. $r_i = v_i^T C_{\underline{X}} v_i$ and 0^T is an n_X row vector of zeros. Notice that the critical values β correspond to the slope of the IB curve, as they represent the Lagrange multipliers of the IB problem.

Unfortunately, this solution is limited to jointly Gaussian random variables. In fact, it can be shown that a closed form analytical solution (for continuous random variables) may only exist under quite restrictive assumptions on the underlaying distribution. Moreover, as the IB curve is so challenging to evaluate in the general case, most known attempts either focus on extending the GIB to other distributions under varying assumptions, or approximate the IB curve by different means.

Rey and Roth (2012) reformulate the IB problem in terms of probabilistic copulas. They show that under a Gaussian copula assumption, an analytical solution (which extends the GIB) applies to joint distributions with arbitrary marginals. This formulation provides several interesting insights on the IB problem. However, its practical implications are quite limited as the Gaussian copula assumption is very restrictive. In fact, it implicitly requires that the joint distribution would maintain a Gaussian structure. As we show in the following sections, this assumption makes the problem significantly easier and does not hold in general.

Chalk et al. (2016) provide a lower bound to the IB curve by using an approximate variational scheme, analogous to variational expectation maximization. Their method relaxes the IB problem by restricting the class of distributions, $P(\underline{Y}|\underline{T})$ and $P(\underline{T})$ to a set of parametric models. This way, the relaxed IB problem may be solved in EM-like steps; their suggested algorithm iteratively maximizes the objective over the mappings (for fixed parameters) and then maximize the set of parameters, for fixed mappings. Chalk et al. (2016) show that this method can be effectively applied to "sparse" data in which $\underline{X}$ and $\underline{Y}$ are generated by sparsely occurring latent features. However, in the general case, their suggested bound strongly depends on the assumption that the chosen parametric models provide reasonable approximations for the optimal distributions. This assumption is obviously quite restrictive. Moreover, it is usually difficult to validate, as the optimal distributions are unknown. Kolchinsky et al. (2017) take a somewhat similar approach, as they suggest a variational upper bound to the IB curve. The main difference between the two methods relies on the variational approximation of objective, $I(\underline{X}; \underline{Y})$. However, they are both prone to the same difficulties stated above.

Alemi et al. (2016) propose an additional variational inference method to construct a lower bound to the IB curve. Here, they re-parameterize the IB problem followed by Monte Carlo sampling, to get an unbiased estimate of the IB objective gradient. This allows them to apply deep neural networks in order to parameterize any given distribution. However,

this method fails to provide guarantees on the obtained bound, as a result of the suggested stochastic gradient descent optimization approach.

Achille and Soatto (2018) relax the bottleneck problem by introducing an additional *total correlation* (TC) regularization term that strives to maximize the independence among the components of the representation T. They show that under the assumption that the Lagrange multipliers of the TC and MI constraints are identical, the relaxed problem may be solved by adding auxiliary variables. However, this assumption is usually invalid, and the suggested method fails to provide guarantees on the difference between the obtained objective and original IB formulation.

In this work we suggest a novel lower bound to the IB curve which provides both theoretical and practical guarantees. In addition, we introduce upper and lower bounds to our suggested solution that are very easy to attain. This way we allow immediate benchmarks to the IB curve using common off-the-shelf methods.

3. Problem formulation

Throughout this manuscript we use the following standard notation: underlines denote vector quantities, where their respective components are written without underlines but with index. For example, the components of the n-dimensional vector $\underline{X}$ are $X_1, X_2, \dots X_n$. Random variables are denoted with capital letters while their realizations are denoted with the respective lower-case letters. The mutual information of two random variables is defined as $I(\underline{X};\underline{Y}) = h(\underline{X}) + h(\underline{Y}) - h(\underline{X},\underline{X})$ where $h(\underline{X}) = -\int_{\underline{X}} f_{\underline{X}}(\underline{x}) \log f_{\underline{X}}(\underline{x}) d\underline{x}$ is the differential entropy of $\underline{X}$ and $f_{\underline{X}}(\underline{x})$ is its probability density function.

We begin by introducing a Gaussian lower bound to the mutual information $I(\underline{X};\underline{Y})$. We then extend our result to the entire IB curve.

3.1 Problem statement

Let $\underline{X} \in \mathbb{R}^{d_x}, \underline{Y} \in \mathbb{R}^{d_y}$ be two multivariate random vectors with a joint cumulative distribution function (CDF) $F_{XY}(x,y)$ and mutual information $I(\underline{X},\underline{Y})$. In the following sections we focus on bounding $I(\underline{X},\underline{Y})$ from below with an analytical expression. Let $\underline{U} = \phi(\underline{X})$ and $\underline{V} = \psi(\underline{Y})$ be two transformations of $\underline{X}$ and $\underline{Y}$, respectively. Assume that $\underline{U} \in \mathbb{R}^{d_u}$ and $\underline{V} \in \mathbb{R}^{d_v}$ are *separately normally distributed*. This means that $\underline{U} \sim N\left(\mu_{\underline{U}}, C_{\underline{U}}\right)$ and $\underline{V} \sim N\left(\mu_{\underline{V}}, C_{\underline{V}}\right)$ but the vector $[\underline{U},\underline{V}]^T$ is not necessarily normally distributed. This allows us to derive the following basic inequality

$$\begin{aligned} I(\underline{X},\underline{Y}) \geq & I(\underline{U},\underline{V}) = h(\underline{U}) + h(\underline{V}) - h(\underline{U},\underline{V}) \geq \\ & h(\underline{U}) + h(\underline{V}) - h(\underline{U}_{jg},\underline{V}_{jg}) = \frac{1}{2}\log\left(\frac{\left|C_{[\underline{U},\underline{V}]}\right|}{|C_{\underline{U}}||C_{\underline{V}}|}\right) \end{aligned} \tag{3}$$

where the first inequality follows from the Data Processing lemma (Cover and Thomas, 2012) and the second inequality follows from $\left[\underline{U}_{jg},\underline{V}_{jg}\right]^T$ being jointly Gaussian (jg) distributed with the same covariance matrix as $[\underline{U},\underline{V}]^T$, $C_{[\underline{U}_{jg},\underline{V}_{jg}]} = C_{[\underline{U},\underline{V}]}$, so that $h(\underline{U}_{jg},\underline{V}_{jg}) \geq h(\underline{U},\underline{V})$ (Cover and Thomas, 2012). Notice that (3) can also be derived from an *information geometry* (IG) view point, as shown by Cardoso (2003).

Equality is attained in (3) iff $I(\underline{X},\underline{Y}) = I(\underline{U},\underline{V})$ (no information is lost in the transformation) and $\underline{U} = \phi(\underline{X})$, $\underline{V} = \psi(\underline{Y})$ are jointly normally distributed. In other words, in order to preserve all the information we must find ϕ and ψ that capture all the mutual information, and at the same time make $\underline{X}$ and $\underline{Y}$ jointly normal. This is obviously a complicated task as ϕ and ψ only operate on $\underline{X}$ and $\underline{Y}$ separately. Therefore, we are interested in maximizing this lower bound as much as possible:

$$\begin{aligned} \max_{\phi,\psi} \quad & \log\left(\frac{\left|C_{[\underline{U},\underline{V}]}\right|}{|C_{\underline{U}}||C_{\underline{V}}|}\right) \\ \text{subject to} \quad & \underline{U} = \phi(\underline{X}) \sim N\left(0, C_{\underline{U}}\right) \\ & \underline{V} = \psi(\underline{Y}) \sim N\left(0, C_{\underline{V}}\right) \end{aligned} \tag{4}$$

where the constraints imply that $\underline{U}$ and $\underline{V}$ are separately normally distributed random vectors with zero means and covariance matrices $C_{\underline{U}}$ and $C_{\underline{V}}$, respectively. In other words, we would like to maximize Cardoso (2003) IG bound by applying two transformations, ϕ and ψ, to the original variables. This would allow us to achieve a tighter result.

Notice that our objective is invariant to the means of $\underline{U},\underline{V}$ so they are chosen to be zero. In addition, it is easy to show that our objective is invariant to linear transformations of $\underline{U},\underline{V}$. This means we can equivalently assume that $C_{\underline{U}}, C_{\underline{V}}$ are identity covariance matrices. As shown by Kay (1992) and others (Klami and Kaski, 2005; Chechik et al., 2005), maximizing the objective of (4) is equivalent to maximizing the canonical correlations, $\text{cov}(U_i, V_i)$. Therefore, our problem may be written as

$$\begin{aligned} \max_{\phi,\psi} \quad & \sum_{i=1}^{k} E\left(U_i V_i\right) \\ \text{subject to} \quad & \underline{U} = \phi(\underline{X}) \sim N\left(0, I\right) \\ & \underline{V} = \psi(\underline{Y}) \sim N\left(0, I\right) \end{aligned} \tag{5}$$

where $k = \min\{d_u, d_v\}$. This problem may also be viewed as a variant of the well-known CCA problem (Hotelling, 1936), where we optimize over nonlinear transformations ϕ and ψ, and impose additional normality constraints. As in CCA, this problem can be solved iteratively by gradually finding the the optimal canonical components in each step (subject to the normality constraint), while maintaining orthogonality with the components that were previously found. For simplicity of the presentation we begin by solving (5) in the univariate (1-D) case. Then, we generalize to the multivariate case. In each of these setups we present a solution to the problem, followed by simpler upper and lower bounds.

4. The univariate case

In the univariate case we assume that $d_x = d_y = k = 1$. We seek ϕ, ψ such that

$$\begin{aligned} \max_{\phi,\psi} \quad & \rho = E(UV) \\ \text{subject to} \quad & U = \phi(X) \sim N(0,1) \\ & V = \psi(Y) \sim N(0,1). \end{aligned} \tag{6}$$

As a first step towards this goal, let us relax our problem by replacing the normality constraint with simpler second order statistics constraints,

$$\begin{aligned} \max_{\phi,\psi} \quad & \rho = E(UV) \\ \text{subject to} \quad & U = \phi(X),\ E(U) = 0,\ E(U^2) = 1 \\ & V = \psi(Y),\ E(V) = 0,\ E(V^2) = 1. \end{aligned} \tag{7}$$

As mentioned above, this problem is a non-linear extension of CCA, which traces back to early work by Lancaster (1963). As this problem is also a relaxed version of our original task (6), it may serve us as an upper bound. This means that the optimum of (7), denoted as ρ_{ub}, necessarily bound from above ρ_*, the optimum of (6).

4.1 Alternation Conditional Expectation (ACE)

Breiman and Friedman (1985) show that the optimal solution to (7) is achieved by a simple alternating conditional expectation procedure, named ACE. Assume that $\psi(Y)$ is fixed, known and satistfies the constraints. Then, we optimize (7) only over ϕ and by Cauchy-Schwarz inequality, we have that

$$E(\phi(X)\psi(Y)) = E_x\left(\phi(X)E(\psi(Y)|X)\right) \leq \sqrt{\text{var}(\phi(X))}\sqrt{\text{var}(E(\psi(Y)|X))}$$

with equality iff $\phi(X) = c \cdot E(\psi(Y)|X)$. Therefore, choosing the constant c to satisfy the unit variance constraint we achieve $\phi(X) = \frac{E(\psi(Y)|X)}{\sqrt{var(E(\psi(Y)|X))}}$. In the same manner we may fix $\phi(X)$ and attain $\psi(Y) = \frac{E(\phi(X)|Y)}{\sqrt{var(E(\phi(X)|Y))}}$. These coupled equations are in fact necessary conditions for the optimality of ϕ and ψ, leading to an alternating procedure in which at each step we fix one transformation and optimize the other. Breiman and Friedman (1985) prove that this procedure convergences to the global optimum using Hilbert space algebra. They show that the transformations ϕ and ψ may be represented in a zero-mean and finite variance Hilbert space, while the conditional expectation projection is linear, closed, and shown to be self-adjoint and compact under mild assumptions. Then, the coupled equations may be formulate as an eigen problem in the Hilbert space, for which there exists a unique and optimal solution.

The following lemma defines a strict connection between the non-linear canonical correlations and the Gaussinized IB problem.

Lemma 1 *Let ρ_{ub} be the solution to (7). If $I(X;Y) > -\log\left(1-\rho_{ub}^2\right)$, then there are no transformations ϕ, ψ such that $U = \phi(X)$ and $V = \psi(Y)$ are jointly normally distributed and preserve all of the mutual information, $I(X;Y)$.*

Proof Let ρ_* be the solution to (6). As mentioned above, $\rho_{ub} \geq \rho_*$. Therefore, $I(X;Y) > -\log\left(1-\rho_{ub}^2\right) > -\log\left(1-\rho_*^2\right)$. This means that the inequality (3) cannot be achieved with equality. Hence, there are no transformations $U = \phi(X)$ and $V = \psi(Y)$ so that U and V are jointly normal and preserve all of the mutual information, $I(X;Y)$. ■

Lemma 1 suggests that if the optimal transformations of the relaxed problem (which can be obtained by ACE) fails to capture all the mutual information between X and Y, then there

are no transformations that can project X and Y onto jointly normal variables without losing information. Moreover, notice that the maximal level of correlation ρ_{ub} cannot be further increased, even if we allow $\underline{U} = \phi(X)$ and $\underline{V} = \phi(Y)$ to reside in greater dimensions. This means that Lemma 1 holds for any $\phi : R^{d_x} \to R^{d_u}$ and $\psi : R^{d_y} \to R^{d_v}$, such that $d_u, d_v \geq 0$.

4.2 Alternating Gaussinized Conditional Expectations (AGCE)

Let us go back to our original problem, which strives to maximize the correlation between U and V, subject to marginal normality constraints (6). Here we follow Breiman and Friedman (1985), and suggest an alternating optimization procedure.

Let us fix $\psi(Y)$ and optimize (6) with respect to $\phi(X)$. As before, we can write the correlation objective as $E(\phi(X)\psi(Y)) = E_x\left(\phi(X)E(\psi(Y)|X)\right)$. Since $E(\phi(X)^2)$ is constrained to be equal to 1 while $E\left(E(\psi(Y)|X)^2\right)$ is fixed, maximizing $E_x\left(\phi(X)E(\psi(Y)|X)\right)$ is equivalent to minimizing $E_x\left(\phi(X) - E(\psi(Y)|X)\right)^2$. For simplicity, denote $\bar{X} \equiv E(\psi(Y)|X)$. Then, our optimization problem can be reformulated as

$$\begin{aligned} \min_{\phi} \quad & E\left(\phi(\bar{X}) - \bar{X}\right)^2 \\ \text{subject to} \quad & \bar{X} \sim F_{\bar{X}} \\ & \phi(\bar{X}) \sim N(0,1) \end{aligned} \tag{8}$$

where $F_{\bar{X}}$ is the (fixed) CDF of $\bar{X} \equiv E(\psi(Y)|X)$. Notice that ϕ is necessarily a function of $\bar{X}$ alone (as opposed to X), for simple optimization considerations. Assuming that $\bar{X}$ and $U = \phi(\bar{X})$ are two separable metric spaces such that any probability measure on $\bar{X}$ (or U) is a Radon measure (i.e. they are Radon spaces), then (8) is simply an optimal transportation problem (Monge, 1781) with a strictly convex cost function (mean square error). We refer to $\phi^*(\bar{X})$ that minimizes (8) as the optimal map.

The optimal transportation problem was presented by Monge (1781) and has generated an important branch of mathematics. The problem originally studied by Monge was the following: assume we are given a pile of sand (in $\mathbb{R}^3$) and a hole that we have to completely fill up with that sand. Clearly the pile and the hole must have the same volume and different ways of moving the sand will give different costs of the operation. Monge wanted to minimize the cost of this operation. Formally, the optimal transportation problem is defined as

$$\inf\left\{\int_{\bar{X}} c(\bar{X}, \phi(\bar{X}))d\mu(\bar{X}) \Big| \phi_*(\mu) = \nu\right\}$$

where μ and ν are the probability measures of $\bar{X}$ and U respectively, $c(\cdot,\cdot)$ is some cost function and $\phi_*(\mu)$ denotes the push forward of μ by the map ϕ. Clearly, (8) is a special case of the optimal transportation problem where the $\mu = F_{\bar{X}}$, ν is a standard normal distribution and the cost function is the euclidean distance between the two.

Assume that $\bar{X} \in \mathbb{R}$ has finite p^{th} moments for $1 \leq p < \infty$ and a strictly continuous CDF, $F_{\bar{X}}$ (that is $\bar{X}$ is a strictly continuous random variable). Then, Rachev and Rüschendorf (1998) show that the optimal map (which minimizes (8)) is exactly $\phi^*(\bar{X}) = \Phi_N^{-1} \circ F_{\bar{X}}(\bar{X})$ where Φ_N^{-1} is the inverse CDF of a standard normal distribution. As shown by Rachev and

Rüschendorf (1998), the optimal map is unique and achieves

$$E\left(\left(\phi^*(\bar{X})-\bar{X}\right)^2\right)=\int_0^1\left(F_{\bar{X}}(s)-\Phi_N(s)\right)^2 ds. \tag{9}$$

Notice that the optimal map may be generalized to the multivariate case, as discussed in the next Section. The solution to the optimal transportation problem is in fact the "optimal projection" of our problem (8). Further, it allows us to quantify how much we lose from imposing the marginal normality constraint, compared with ACE's optimal projection.

Notice that the optimal map, $\phi^*(\bar{X})=\Phi_N^{-1}\circ F_{\bar{X}}(\bar{X})$, is simply marginal Gaussianization of $\bar{X}$: applying $\bar{X}$'s CDF to itself results in a uniformly distributed random variable, while Φ_N^{-1} shapes this uniform distribution into a standard normal. In other words, while the optimal projection of $\psi(Y)$ on X is its conditional expectation, the optimal projection under a normality constraint is simply a Gaussianization of the conditional expectation. The uniqueness of the optimal map leads to the following necessary conditions for an optimal solution to (6),

$$\begin{aligned}\phi(X)&=\Phi_N^{-1}\circ F_{E(\psi(Y)|X)}(E(\psi(Y)|X))\\ \psi(Y)&=\Phi_N^{-1}\circ F_{E(\phi(X)|Y)}(E(\phi(X)|Y)).\end{aligned} \tag{10}$$

As in ACE, these necessary conditions imply an alternating projection algorithm, namely, the Alternating Gaussinized Conditional Expectation (AGCE). Here, we begin by randomly choosing a transformation that only satisfies the normality constraint $\psi(Y)\sim N(0,1)$. Then, we iterate by fixing one of the transformation while optimizing the other, according to (10). We terminate once $E(\phi(X)\psi(Y))$ fails to increase, which means that we converged to a set of transformations that satisfy the necessary conditions for optimal solution. Algorithm 1 summarizes our suggested approach. Notice that in every step of our procedure, we may either:

1. Increase our objective value, as a result of the optimal map for (8).
2. Maintain with the same objective value and with the same transformation that was found in of the previous iteration, as we converged to (10).

This means that our alternating method generates a monotonically increasing sequence of objective values. Moreover, as written in Section 4, this sequence is bounded from above by the optimal correlation given by ACE. Therefore, according to the monotone convergence theorem, our suggested method converges to a local optimum.

Unfortunately, as opposed to ACE, our projection operator is not linear and we cannot claim for global optimality. We see that for different random initializations we converge to (a limited number) of local optima. Yet, AGCE provides an effective tool for finding local maximizers of (4), which together with MCMC (Gilks, 2005) initializations (or any other random search mechanisms) is capable of finding the global optimum.

4.3 Off-the-shelf lower bound

Although the AGCE method provides a (locally) optimal solution to (4), we would still like to consider a simpler "off-the-shelf" mechanism that is easier to implement and gives a lower

Algorithm 1 Alternating Gaussinized Conditional Expectations (AGCE) for the univariate case

Require: F_{XY}, the joint distribution function of X and Y.
Require: $g : \mathbb{R} \to \mathbb{R}$, a random mapping.
Set $\psi(Y) = \Phi_N^{-1} \circ F_{g(Y)}(g(Y))$.
Set $\phi(X) = \Phi_N^{-1} \circ F_{E(\psi(Y)|X)}(E(\psi(Y)|X))$.
Set $\rho = E(\phi(X)\psi(Y))$.
Set $T = 0$
while $T \neq 1$ **do**
  Set $\psi(Y) = \Phi_N^{-1} \circ F_{E(\phi(X)|Y)}(E(\phi(X)|Y))$.
  Set $\phi(X) = \Phi_N^{-1} \circ F_{E(\psi(Y)|X)}(E(\psi(Y)|X))$.
  if $E(\phi(X)\psi(Y)) \not> \rho$ **then**
    $T = 1$
  else
    $\rho = E(\phi(X)\psi(Y))$
  end if
end while
return $\phi(X), \psi(Y), \rho$

bound to the best we can hope for. Here, we tackle (4) in two phases. In the first phase we would like to maximize the correlation objective, $E(UV)$, subject to the relaxed second order statistics constraints (as defined in (7)). Then, we enforce the marginal normality constraints by simply applying *separate Gaussianization* to the outcome of the first phase. In other words, we first apply ACE to increase our objective as much as possible, and then separately Gaussianize the results to meet the normality constraints, hoping this process does not reduce our objective "too much". Notice that in this univariate case, separate Gaussianization is achieved according to Theorem 2:

Theorem 2 *Let X be any random variable $X \sim F_X(x)$ and $\theta \sim Unif[0,1]$ be statistically independent of it. In order to shape X to a normal distribution the following applies:*

1. *Assume X is a non-atomic distribution ($F_X(x)$ is strictly increasing) then*

$$\Phi_N^{-1} \circ F_X(X) \sim N(0,1)$$

2. *Assume X is discrete or a mixture probability distribution then*

$$\Phi_N^{-1} \circ (F_X(X) - \theta P_X(x)) \sim N(0,1)$$

The proof of this theorem can be located in Appendix 1 of (Shayevitz and Feder, 2011). Theorem 2 implies that if X is strictly continuous then we may achieve a normal distribution by applying $\Phi_N^{-1} \circ F_X(X)$ to it, as discussed in the previous section. Otherwise, we shall handle its CDF's singularity points by randomly scattering them in a uniform manner, followed by applying Φ_N^{-1} to the random variable we achieved. Notice that this process do not allow any flexibility in the Gaussianization process. However, we show that in the

multivariate case (Section (5.3)) the equivalent process is quite flexible and allows us to control the correlation objective.

Further, notice that this lower bound is by no means a candidate for an optimal solution to (6), as it does not meet the necessary conditions described in (10). Yet, by finding both an upper and lower bounds (through ACE, and then separately Gaussianizing the result of ACE) we may immediately achieve the range in which the optimal solution necessarily resides. Assuming this range is not too large, one may settle for a sub-optimal solution without a need to apply AGCE at all.

4.4 Illustrative example

We now demonstrate our suggested methodology with a simple illustrative example. Let $X \sim N(0,1)$, $W \sim N(0,\epsilon^2)$ and $Z \sim N(\mu_z,1)$ be three normally distributed random variables, all independent of each other. Let P be a Bernoulli distributed random variable with a parameter $\frac{1}{2}$, independent of X, W and Z. Define Y as:

$$Y = \left\{ \begin{array}{ll} \text{X+W} & \text{P=0} \\ \text{Z} & \text{P=1} \end{array} \right\}.$$

Then, Y is a balanced Gaussian mixture with parameters

$$\theta_y = \left\{\mu_1 = 0, \sigma_1^2 = 1+\epsilon^2, \mu_2 = \mu_z, \sigma_2^2 = 1\right\}.$$

The joint probability density function of X and Y is also a balanced two-dimensional Gaussian mixture with parameters

$$\theta_{xy} = \left\{\mu_1 = \left[\begin{array}{c} 0 \\ 0 \end{array}\right], C_1 = \left[\begin{array}{cc} 1 & 1 \\ 1 & 1+\epsilon^2 \end{array}\right], \mu_2 = \left[\begin{array}{c} 0 \\ \mu_z \end{array}\right], C_2 = I\right\}.$$

Let us further assume that μ_z is large enough, and ϵ^2 is small enough, so that the overlap between the two Gaussian is negligible. For example, we set $\mu_z = 10$ and $\epsilon = 0.1$. The correlation between X and Y is easily shown to be $\rho_{xy} = \frac{1/2}{\sqrt{1+1/2\epsilon^2+1/4\mu_z}} = 0.098$. The mutual information between X and Y is defined as

$$I(X;Y) = h(X) + h(Y) - h(X;Y).$$

Since we assume that the Gaussians in the mixture practically do not overlap, we have that

$$h(Y) = -\int f_Y(y)\log f_Y(y)dy \approx \frac{1}{4}\log\left(2\pi e(1+\epsilon^2)\right) + \frac{1}{4}\log\left(2\pi e\right) + 1. \tag{11}$$

In the same manner,

$$\begin{aligned} h(X,Y) = -\int f_{X,Y}(x,y)\log f_{X,Y}(x,y)dxdy \approx \\ \frac{1}{4}\log\left((2\pi e)^2|C_1|\right) + \frac{1}{4}\log\left((2\pi e)^2|C_2|\right) + 1. \end{aligned} \tag{12}$$

Plugging $\mu_z = 10$ and $\epsilon = 0.1$ we have that

$$I(X;Y) = h(X) + h(Y) - h(X;Y) \approx 1.66\text{bits}. \tag{13}$$

The scatter plot on the left of Figure 1 illustrates $10,000$ independent draws of X and Y, where the blue circles corresponds to the "correlated samples" ($P = 0$) while the blue crosses are the "noise" ($P = 1$).

Before we proceed to apply our suggested methods, let us first examine two benchmark options for separate Gaussianization. As an immediate option, we may always apply separate Gaussianization, directly to X and Y, denoted as U_a and V_a respectively. This corresponds to Cardoso (2003) information geometry bound. Since X is already normally distributed we may set $U_a = X$ and only apply Gaussinization to Y. Let $V_a = \psi(Y)$ be the Gaussianization of Y. This means that

$$V_a = \Phi_N^{-1}\left(F_Y\left(Y\right)\right) = \Phi_N^{-1}\left(\Phi_{GM(\theta_y)}(Y)\right)$$

where $\Phi_{GM(\theta_y)}$ is the cumulative distribution function a Gaussian Mixture with the parameters θ_y described above. Therefore,

$$\rho_{u_a,v_a} = E(XV) = \frac{1}{2}E\left(X\Phi_N^{-1}\left(\Phi_{GM(\theta_y)}(X+W)\right)\right).$$

Although it is not possible to obtain a closed form solution to this expectation, it may be numerically evaluated quite easily, as X and W are independent. Assuming $\mu_z = 10$ and $\epsilon = 0.1$ we get that $\rho_{u_a,v_a} \approx 0.288$ and our lower bound on the mutual information, as appears in (3), is $I_g \equiv -\frac{1}{2}\log\left(1-\rho_{u_a,v_a}^2\right) \approx 0.0628$bits. The middle scatter plot of Figure 1 presents this separate marginal Gaussianization of the previously drawn $10,000$ samples of X and Y. Notice that the marginal Gaussianization is a monotonic transformation, so that the Y samples are not being shuffled and maintain the separation between the two parts of the mixture. While the red circles are now "half Gaussian", the blue crosses are shaped in a curvy manner, so that their marginal distribution (projected on the y axis) is also a "half Gaussian", leading to a normal marginal distribution of Y. We notice that while the mutual information between X and Y is 1.66 bits, the lower bound attained by this naive Gaussianization approach is close to zero. This is obviously an unsatisfactory result.

A second benchmark alternative for separate Gaussianization is to take advantage of the Gaussian mixture properties. Since we assume that the two Gaussians of Y are practically separable, we may distinguish between observations from the two Gaussians. Therefore, we can simply reduce μ_z from the Z samples (the red circles), and normalize the observations of $X + W$. This way the transformed Y becomes a Gaussian mixture of two co-centered standard Gaussians, and no further Gaussianization is necessary. For $\mu_z = 10$ and $\epsilon = 0.1$, this leads to a correlation of

$$\rho_{u_b,v_b} = \frac{1}{2}E\left(\frac{1}{\sqrt{1+\epsilon^2}}(X+W)X\right) = \frac{1}{2}\frac{1}{\sqrt{1+\epsilon^2}} = 0.497 \tag{14}$$

and a corresponding mutual information lower bound of $I_g = 0.204$ bits. However, notice that the suggested transformation is not invertible and may cause a reduction in mutual information. Specifically, we now have that the joint distribution of $U_b = X$ and V_b follows a Gaussian mixture model with parameters:

$$\theta_{u_b,v_b} = \left\{\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, C_1 = \begin{bmatrix} 1 & \frac{1}{\sqrt{1+\epsilon^2}} \\ \frac{1}{\sqrt{1+\epsilon^2}} & 1 \end{bmatrix}, \mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, C_2 = I\right\}.$$

Therefore,

$$h(U_b, V_b) = -\int f_{U_b,V_b}(u,v)\log\left(f_{U_b,V_b}(u,v)\right)dudv = \tag{15}$$
$$-\int \phi_{GN(\theta_{u_b,v_b})}(u,v)\log\phi_{GN(\theta_{u_b,v_b})}(u,v)dudv \approx 3.1384\text{bits}$$

where $\phi_{GN(\theta_{u_b,v_b})}(u,v)$ is the probability density function of a Gaussian mixture with the parameters θ_{u_b,v_b} described above, and the last approximation step is due to numerical integration. This leads to $I(U_b;V_b) = 0.95$ bits.

To conclude, although the mutual information is reduced from 1.66 bits to 0.95 bits, the suggested bound increased quite dramatically, from 0.0628 bits to 0.204 bits. The right plot of Figure 1 demonstrates this customized separate Gaussianization (as it only applies for this specific setup) to the previously sampled X and Y. Again, we emphasis that this solution is not applicable in general, and is only feasible due to the specific nature of this Gaussian mixture model.

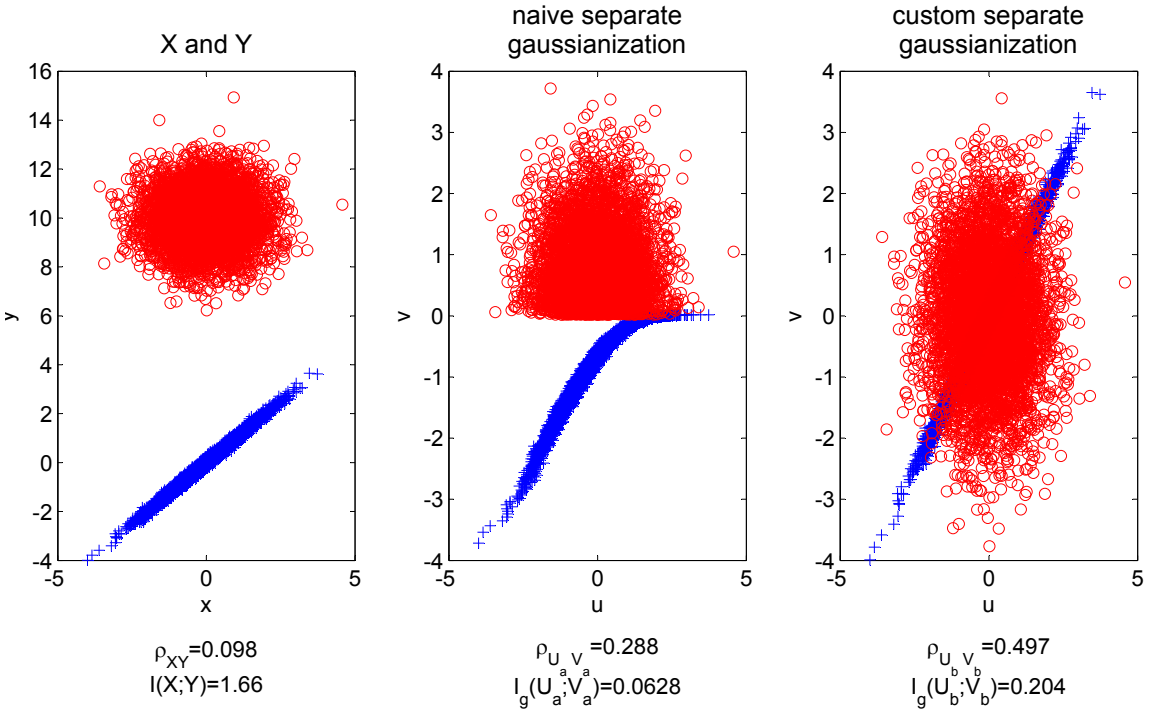

Figure 1: Univariate Gaussianization: Left: scatter of X and Y. Middle: naive separate Gaussianization to X and Y. Right: separate Gaussianization which considers the Gaussian Mixture model of X and Y, as described in the text.

Let us now turn to our suggested methods, as described in detail in the previous sections. We begin by applying the ACE procedure (Section 4.1), to attain an upper bound on our problem (6). Not surprisingly, ACE converges to a solution in which the samples of Y that are independent of X (the ones that come from Z) are set to zero,

while the rest are normalized to achieve an unit variance. Therefore, the resulting correlation is $\rho_{ub} = \frac{1/2}{\sqrt{1/2(1+\epsilon^2)}} = 0.703$. This results further implies that we can never find a Gaussianization procedure that will capture all the information between X and Y, as $I(X;Y) > -\log\left(1-\rho_{ub}^2\right) = 0.4917$ bits, according to Lemma 1. The left scatter plot of Figure 2 demonstrates the outcome of the ACE procedure, applied to the drawn $10,000$ samples of X and Y.

Next, we apply our suggested AGCE routine, described in Section 4.2. As discussed above, the AGCE only converges to a local optimum. Therefore, we initialize it with several random transformation (including the ACE solution that we just found). We notice that the number of convergence points is very limited and results in almost similar maxima. The middle scatter plot of Figure 4.2 shows the best result we achieve, leading to a correlation coefficient of 0.66 and a lower bound on a corresponding Gaussian lower bound (3) of 0.411 bits. This result demonstrates the power of our suggested approach, as it significantly improves the benchmarks, even compared with the U_b, V_b that considers the separable Gaussian mixture nature of our samples.

Finally, we evaluate a lower bound for (6), as described in Section 4.3. Here, we simply apply separate Gaussianization to the outcome of the ACE procedure. This results in $\rho_{lb} = 0.646$ and a corresponding $I_g = 0.389$. The right scatter plot of Figure 2 shows the Gaussianized samples the we achieve. We notice that this lower bound is not significantly lower than AGCE, suggesting that in some cases we may settle for this less involved method.

To conclude, our suggested solution surpasses the benchmarks quite easily, as we increase the lower bound from 0.204 bits using the custom Gaussianization procedure to 0.411 bits using our suggested solution. We notice that all of the discussed procedures result in joint distributions that are quite far from normal. This is not surprising, since X and Y were highly "non-normal" to begin with. Specifically, all of the suggested procedures lose information, compared with the original $I(X;Y) = 1.66$. However, our suggested solution minimizes this loss, and may be considered "more jointly normal" than others, in this regards.

5. The multivariate case

Let us now consider the multivariate case where both $\underline{X} \in \mathbb{R}^{d_x}$ and $\underline{Y} \in \mathbb{R}^{d_y}$ are random vectors with a joint CDF $F_{\underline{X},\underline{Y}}$. One of the fundamental differences from the univariate case is that Gaussianizing each of these vectors (even separately) is not a simple task. In other words, finding a transformation $\phi : \mathbb{R}^{d_x} \to \mathbb{R}^{d_u}$ such that $\underline{U} = \phi(\underline{X})$ is normally distributed may be theoretically straight-forward but practically involved.

For the simplicity of the presentation, assume that $\underline{X} = [X_1, X_2]^T$ is a two dimensional, strictly continuous, random vector. Then, Gaussianization may be achieved in two steps: first, apply marginal Gaussianization to X_1, so that $U_1 = \Phi_N^{-1} \circ F_{X_1}(X_1)$. Then, apply marginal Gaussianization on X_2, conditioned on each possible realization of the previous component, $U_2|u_1 = \Phi_N^{-1} \circ F_{X_2|U_1}(X_2|U_1 = u_1)$. This results in a jointly normally distributed vector $\underline{U} = [U_1, U_2]^T$. While this procedure is theoretically simple, it is quite problematic to apply in practice, as it requires Gaussianizing each and every conditional CDF. This is obviously impossible, given a finite number of samples. Yet, it gives us a constructive

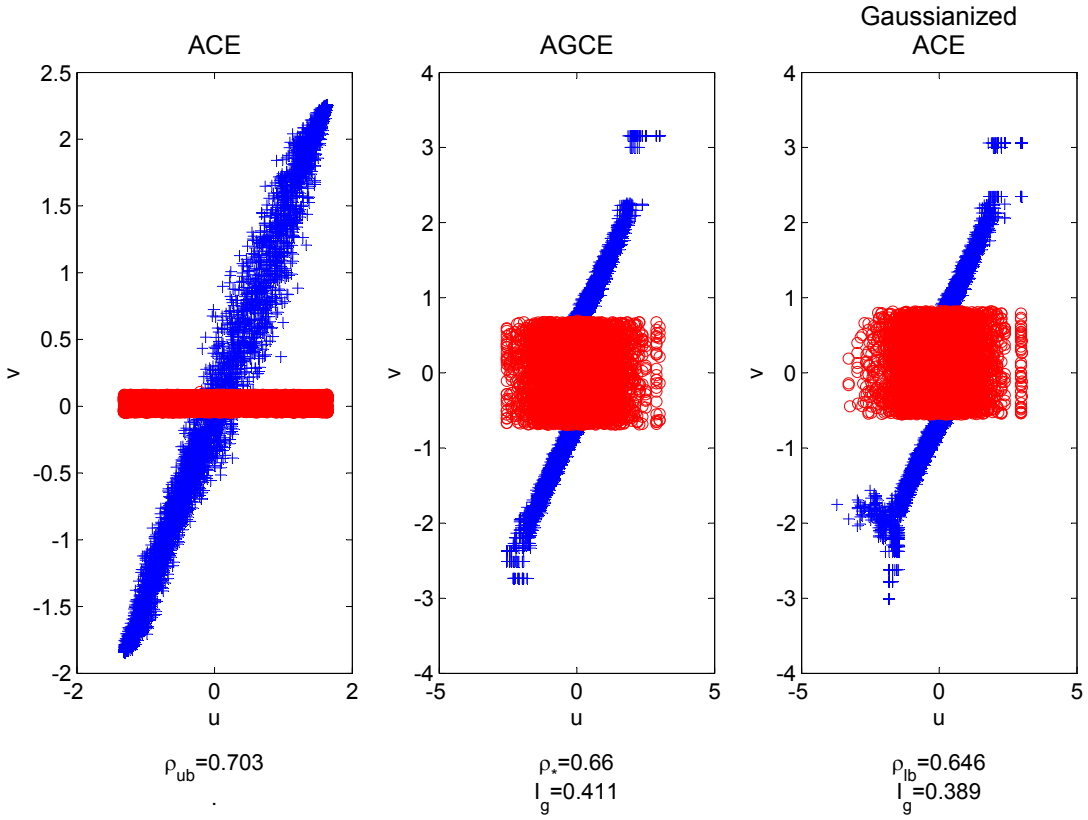

Figure 2: Our Suggested Univariate Gaussianization Schemes: Left: upper bound by ACE. Middle: (local) optimal solution by AGCE. Right: lower bound by separate Guassianization to ACE.

method, assuming that all the CDF's are known. In the following sections we shall present several alternatives for Gaussianization in a finite sample size setup.

5.1 Upper bound by ACE

As in the univariate case, we begin our analysis by relaxing the normality constraints with softer second order statistics constraints. This leads to an immediate multivariate generalization of the ACE procedure:

We begin by extracting the first canonical pair, which satisfies $U_1 = c \cdot E(V_1|\underline{X})$ and $V_1 = c \cdot E(U_1|\underline{Y})$. As in the univariate case, c is a normalization coefficient (the square root of the variance of the conditional expectation), and the optimization is done by alternating projections. Then, we shall extract the second pair of canonical components, subject to an orthogonality constraint with the first pair. It is easy to show that if V_2 is orthogonal to V_1, then $U_2 = c \cdot E(V_2|\underline{X})$ is orthogonal to U_1, and obviously maximizes the correlation with V_2. Therefore, we may extract the second canonical pair by first randomly assigning a zero-mean and unit variance V_2 that is also orthogonal to V_1 (by the Gram-Schmidt procedure, for example), followed by alternating conditional expectations with respect to V_2 and U_2, in the same manner as we did with the first pair. We continue this way for the rest of the

canonical pairs. As in the univariate case, convergence to a global maximum is guaranteed from the same Hilbert space arguments. As before, the multivariate ACE sets an upper bound to (5) as it maximizes a relaxed version of this problem.

Lemma 3 *Let $\underline{U}_*, \underline{V}_*$ be the outcome of multivariate ACE procedure (the canonical vectors). Assuming that $I(\underline{X};\underline{Y}) > \log\left|C_{[\underline{U}_*,\underline{V}_*]}\right|$, then there are no transformations such that $\underline{U} = \phi(\underline{X})$ and $\underline{V} = \psi(\underline{Y})$ follow a jointly normal distribution and preserve all of the mutual information, $I(X;Y)$.*

The proof of Lemma 3 follows exactly from the proof of Lemma 1. Here again, the multivariate ACE objective, $\log\left|C_{[\underline{U}_*,\underline{V}_*]}\right|$, cannot be further increased by artificially inflating the dimension of the problem. Therefore, Lemma 3 holds for any $\phi : \mathbb{R}^{d_x} \to \mathbb{R}^{d_u}$ and $\psi : \mathbb{R}^{d_y} \to \mathbb{R}^{d_v}$, such that $d_u, d_v \geq 0$.

5.2 multivariate AGCE

As with the multivariate ACE, we propose a generalized multivariate procedure for AGCE. We begin by extracting the first pair, in the same manner as we did in the univariate case. That is, we find a pair U_1 and V_1 that satisfies

$$\begin{aligned} U_1 &= \Phi_N^{-1} \circ F_{E(U_1|\underline{X})}(E(U_1|\underline{X})) \\ V_1 &= \Phi_N^{-1} \circ F_{E(V_1|\underline{Y})}(E(V_1|\underline{Y})) \end{aligned} \tag{16}$$

by applying the alternating optimization scheme. As we proceed to the second pair, we require that U_2 is both orthogonal and jointly normally distributed with U_1 (same goes for V_2 with respect to V_1). This means that the second pair needs not only to be orthogonal, but also statistically independent with the first pair. In other words, assuming V_2 is fixed, our basic projection step is

$$\begin{aligned} \max_{\phi_2} \quad & E\left(\phi_2(\underline{X})V_2\right) \\ \text{subject to} \quad & \phi_2(\underline{X}) \sim N(0,1) \\ & \phi_2(\underline{X}) \perp\!\!\!\perp \phi_1(\underline{X}). \end{aligned} \tag{17}$$

Let us denote a subspace $\tilde{\underline{X}} \subset \underline{X}$ that is statistically independent of $U_1 = \phi_1(\underline{X})$. Then, the problem of maximizing $E\left(\phi_2(\tilde{\underline{X}})V_2\right)$ subject to $\phi_2(\tilde{\underline{X}}) \sim N(0,1)$ is again solved by the optimal map, $\phi_2(\tilde{\underline{X}}) = \Phi_N^{-1} \circ F_{E(V_2|\tilde{\underline{X}})}(E(V_2|\tilde{\underline{X}}))$. Therefore, the remaining task is to find the "best" subspace $\tilde{\underline{X}} \subset \underline{X}$, so that $E\left(\phi_2(\tilde{\underline{X}})V_2\right)$ is maximal, when plugging the optimal map.

Proposition 4 *Let $U_1 = u_1$ be the value (realization) of U_1. Let $\tilde{\underline{X}} = g\left(\underline{X}, u_1\right)$ be a subspace of $\underline{X}$, independent of U_1. If $g\left(\underline{X}, u_1\right)$ is an invertible function with respect to $\underline{X}$ given u_1, then $\tilde{\underline{X}}$ is an optimal subspace for maximizing $E\left(\phi_2(\tilde{\underline{X}})V_2\right)$ subject to $\phi_2(\tilde{\underline{X}}) \sim N(0,1)$.*

Proof Assume there exists a different subspace $\underline{\tilde{X}}' = g'\left(\underline{X}, u_1\right)$ so that

$$\max_{\phi_2'} E\left(\phi_2'(\underline{\tilde{X}}')V_2\right) > \max_{\phi_2} E\left(\phi_2(\underline{\tilde{X}})V_2\right)$$

subject to the normality constraint. Since g is invertible we have that $\underline{X} = g^{-1}(\underline{\tilde{X}}, u_1)$. Therefore, $\underline{\tilde{X}}' = g'\left(g^{-1}(\underline{\tilde{X}}, u_1)\right) \equiv f(\underline{\tilde{X}}, u_1)$. Plugging this to the inequality above leads to

$$\max_{\phi_2'} E\left(\phi_2'(f(\underline{\tilde{X}}, u_1))V_2\right) > \max_{\phi_2} E\left(\phi_2(\underline{\tilde{X}})V_2\right)$$

which obviously contradicts the optimality of maximization over ϕ_2. ■

Therefore, we are left with finding $\underline{\tilde{X}} = g\left(\underline{X}, u_1\right)$ that is a subspace of $\underline{X}$, independent of U_1 and invertible with respect to $\underline{X}$ given u_1. For simplicity of the presentation, let us first assume that X is univariate. Then, the function $g\left(X, u_1\right) = F_{X|U_1}(X|U_1 = u1)$ is independent of U_1 (as it holds the same (uniform) distribution, regardless to the value of U_1), and invertible given u_1 (assuming that the conditional CDF's $F_{X|U_1}(X|U_1 = u1)$ are continuous for every u_1). Going back to the multivariate $\underline{X} \in \mathbb{R}^{d_x}$, we may follow the same rationale by choosing a single d_x-dimensional distribution that all the conditional CDF's, $F_{\underline{X}|U_1}$ will be shaped to. For simplicity we choose a d_x-dimensional uniform distribution, denoted by its CDF as F_{unif}. Then, $g_*\left(F_{\underline{X}|U_1}, u_1\right) = F_{unif}$, where $g_*(P, x) = Q$ refers to a mapping that pushes forward the distribution P into Q, given x. Specifically, if $p(w)$ and $q(w)$ are the corresponding density functions of the (absolutely continuous) CDF's P and Q respectively, then we know from basic probability theory that the push forward transformation S satisfies

$$p(w) = q\left(S(w)\right)\left|J_S\left(S(w)\right)\right|$$

where J_S is the Jacobian operator of the map S.

To conclude, in order to construct $\underline{\tilde{X}}$ that is independent of U_1 and invertible given u_1, we need to push forward all the conditional CDF's $F_{\underline{X}|U_1}(\underline{X}|U_1 = u_1)$ into a predefined distribution (say, uniform). Then, the optimal map $\phi_2(\underline{\tilde{X}})$ that maximizes $E\left(\phi_2(\underline{\tilde{X}})V_2\right)$ subject to $\phi_2(\underline{\tilde{X}}) \sim N(0,1)$ is given by $\phi_2(\underline{\tilde{X}}) = \Phi_N^{-1} \circ F_{E(V_2|\underline{\tilde{X}})}(E(V_2|\underline{\tilde{X}}))$. In the same manner, we may find $\underline{\tilde{Y}}$ that is independent of V_1 and invertible given v_1, and carry on with the alternating projections. This process continues for all the Gaussinized canonical components and converges to a local optimum, from the same considerations described in the univariate case.

It is important to notice that while this procedure may be considered practically infeasible (as it requires estimating the conditional CDF's), it is equivalently impractical as the multivariate Gaussianization considered in the beginning of this section. Yet, it gives us a local optimum for our problem, assuming that the joint probability distribution is known.

5.3 Off-the-shelf lower bound in the multivariate case

In the same manner as with the univariate case, we may apply a simple off-the-shelf lower bound to (4) by first maximizing the objective as much as we can (using multivariate ACE) followed by Gaussianizing the outcome vectors, hoping we do not reduce the objective

"too much". However, as mentioned in the beginning of Section 5, applying multivariate Gaussianization may be practically infeasible. Therefore, we begin this section by reviewing practical multivariate Gaussianization methodologies. Then, we use these ideas to suggest a practical lower bound, which unlike the univariate case, is not oblivious to our objective.

5.3.1 Practical multivariate Gaussianization

The Gaussianization procedure strives to find a transformation $\underline{Z} = \mathcal{G}(\underline{X})$ so that $\underline{Z} \sim N(0, I)$. A reasonable a cost function for describing "how Gaussian" $\underline{Z}$ really is, may be the Kullback Leibler divergence (KLD) between $\underline{Z}$'s PDF, $f_{\underline{Z}}(\underline{z})$, and a standard normal distribution,

$$J(\underline{Z}) = D_{KL}\left(f_{\underline{Z}}(\underline{z})||f_N(\underline{Z})\right) = \int_{\underline{Z}} f_{\underline{Z}}(\underline{z}) \log\left(\frac{f_{\underline{Z}}(\underline{z})}{f_N(\underline{Z})}\right) dz$$

where $f_N(\underline{Z})$ is the PDF of a standard normal distribution. As shown by Chen and Gopinath (2001), $J(\underline{Z})$ may be decomposed into

$$J(\underline{Z}) = D_{KL}\left(f_{\underline{Z}}(\underline{z})||\prod_{i=1}^{d_z} f_{Z_i}(z_i)\right) + \sum_{i=1}^{d_z} D_{KL}\left(f_{Z_i}(z_i)||f_N(z_i)\right) \tag{18}$$

where the first term quantifies how independent are the components of $\underline{Z}$, and the second term indicates how normally distributed they are. This decomposition led Chen and Gopinath (2001) to an iterative algorithm. In each iteration, their suggested approach applies Independent Component Analysis (Hyvärinen et al., 2004), to minimize the first term, followed by marginal Gaussianization of each component (as we describe for the univariate case), to minimize the second term. Chen and Gopinath show that minimizing one term does not effect the other, which leads to a monotonically decreasing procedure that converges once $\underline{Z}$ is normally distributed.

Notice that the Independent Component Analysis (ICA) is a linear operator. Therefore, if $\underline{Z}$ can be linearly decomposed into independent components, then Chen and Gopinath's Gaussianizion process converges in a single step. Moreover, notice that this Gaussianization process does not require estimating the multivariate distribution. However, it does require estimating the marginals, f_{Z_i} which is considered a much easier task, in general.

A similar yet different multivariate Gaussianization approach was suggested by Laparra et al. (2011). Here, the authors propose to replace the computationally costly ICA with a simple random rotation matrix. This way, they abandon the effort of minimizing the first term of (18), and only shuffle the components so that consequent marginal Gaussianization would further decrease $J(\underline{Z})$. Although this approach takes more iterations to converge to a normal distribution (as in each iteration, only the second term of (18) is being minimized), it holds several favorable properties. First, the overall run-time is dramatically shorter, since applying random rotations is much faster then applying linear ICA. Second, it implies a degree of freedom in choosing the rotation matrix, as the suggested random matrix is just one example of linear shuffling of the components.

5.3.2 Bi-terminal multivariate Gaussianization

Going back to our problem, we would like to Gaussinize $\underline{U}_*$ and $\underline{V}_*$, the outcomes of the multivariate ACE procedure described above. Ideally, we would like to do so while refraining

(as much as we can) from reducing our objective,

$$\log\left(\frac{\left|C_{[\underline{U}_*,\underline{V}_*]}\right|}{|C_{\underline{U}_*}||C_{\underline{V}_*}|}\right). \tag{19}$$

Following the Gaussianization procedures described in the previous section, we suggest an iterative process, where in each iteration we apply a rotation matrix to both vectors, followed by marginal Gaussianization to each of the components of the two vectors. It is easy to show that (19) is invariant to full rank linear transformations. However, it may be effected by the (non-linear) marginal Gaussianization of the components (as described in Theorem 2). Therefore, we would like to find rotation matrices that minimize the effect of the consequent marginal Gaussianization step. This problem is far from trivial. In fact, due to the complicated nature of the marginal Gaussianization procedure, it is quite impossible to minimize the effect of the marginal Gaussianization a-priori, without actually applying it and see how it effects (19). Therefore, we suggest a stochastic search mechanism, which allows us to construct a "reasonable" rotation matrix.

Our suggested mechanism works as follow: At each iteration we begin by drawing two random rotation matrices R_1 and R_2 for the two vectors we are to Gaussianize, just like Laparra et al. (2011). We apply marginal Gaussianization to all the components and evaluate our objective (19). Then, we randomly choose two dimensions and an angle, θ, and construct a corresponding rotation matrix $\tilde{R}$ that rotates the space spanned by the two dimensions in θ degrees. We apply $\tilde{R}\cdot R_1$ to our vector, followed by marginal Gaussianization, and again evaluate (19). If the objective increases we assign $R_1 = \tilde{R}\cdot R_1$. We repeat this process a configurable number of times, for the two vectors we are to Gaussianize.

Notice that our suggested procedure applies a stochastic hill climbing (SHC) search in each step: it randomly searches for the best rotation matrix by gradually composing "small" rotation steps (of two dimensions and an angle), as the complete search space is practically infinite. This procedure guarantees to converge to two multivariate normal vectors, as shown by Laparra et al. (2011), under the reasonable assumption that R_1 and R_2 do not repeatedly converge to identity matrices. Our suggested approach is described in detail in Algorithm 2.

As we see in our experiments, the Bi-terminal Gaussianization process is superior to naively applying a Gaussianization procedure to each of the vectors separately (as suggested by Chen and Gopinath (2001) or Laparra et al. (2011)), in all the cases we examine.

5.4 Illustrative examples

We now examine our suggests multivariate approach in different setups. As in the univariate case, we draw samples from a given model and bound from below the mutual information $I(\underline{X},\underline{Y})$ according to (3). First, we apply the multivariate ACE procedure (Section 5.1) to achieve an upper bound to our objective. Then, we apply separate Gaussianization to ACE's outcome, to attain an immediate lower bound (Section 5.3.1). Further, we tighten this lower bound by replacing the separate Gaussianization with bi-terminal Gaussianization to ACE's outcome (Section 5.3.2). Since our multivariate AGCE procedure (Section 5.2) is practically infeasible, we refrain from using it. This would be further justified later in our results, as we see that the gap between the lower and upper bounds is relatively small. In

Algorithm 2 Bi-terminal multivariate Gaussianization

Require: $\underline{X} \in \mathbb{R}^{d_x}$, $\underline{Y} \in \mathbb{R}^{d_y}$.
Require: Th, a Gaussianization convergence threshold and N, the SHC parameter.
Set $\underline{U} = \underline{X}$ and $\underline{V} = \underline{Y}$.
Set $J_U = J(\underline{U})$ and $J_V = J(\underline{V})$ according to (18).
while $J_U \geq Th$ OR $J_V \geq Th$ **do**
  Draw a rotation martix R_1 of dimensions $d_x \times d_x$ and set $\underline{U}^* = R_1\underline{U}$.
  Draw a rotation martix R_2 of dimensions $d_y \times d_y$ and set $\underline{V}^* = R_2\underline{V}$.
  Apply marginal Gaussianization to $\underline{U}^*$ and $\underline{V}^*$.
  Set $\rho^* = \log\left(\frac{\left|C_{[\underline{U}^*,\underline{V}^*]}\right|}{|C_{\underline{U}^*}||C_{\underline{V}^*}|}\right)$.
  for all $n = 1$ to N **do**
    Draw an angle θ.
    Draw (without replacement) two dimensions d_a, d_b from the set $\{1, \ldots, d_x\}$.
    Construct a rotation matrix $\tilde{R}$ from θ, d_a, d_b and set $\tilde{\underline{U}} = \tilde{R}R_1\underline{U}$.
    Apply marginal Gaussianization to $\tilde{\underline{U}}$.
    Set $\tilde{\rho} = \log\left(\frac{\left|C_{[\tilde{\underline{U}},\underline{V}^*]}\right|}{|C_{\tilde{\underline{U}}}||C_{\underline{V}^*}|}\right)$.
    if $\tilde{\rho} > \rho^*$ **then**
      Set $R_1 = \tilde{R}R_1$, $\underline{U}^* = \tilde{\underline{U}}$ and $\rho^* = \tilde{\rho}$.
    end if
    Draw an angle θ.
    Draw (without replacement) two dimensions d_a, d_b from the set $\{1, \ldots, d_y\}$.
    Construct a rotation matrix $\tilde{R}$ from θ, d_a, d_b and set $\tilde{\underline{V}} = \tilde{R}R_2\underline{U}$.
    Apply marginal Gaussianization to $\tilde{\underline{V}}$.
    Set $\tilde{\rho} = \log\left(\frac{\left|C_{[\underline{U}^*,\tilde{\underline{V}}]}\right|}{|C_{\underline{U}^*}||C_{\tilde{\underline{V}}}|}\right)$.
    if $\tilde{\rho} > \rho$ **then**
      Set $R_2 = \tilde{R}R_2$, $\underline{V}^* = \tilde{\underline{V}}$ and $\rho^* = \tilde{\rho}$.
    end if
  end for
  Set $\underline{U} = \underline{U}^*$ and $\underline{V} = \underline{V}^*$.
  Set $J_U = J(\underline{U})$ and $J_V = J(\underline{V})$ according to (18).
end while
return $\underline{U}, \underline{V}, \rho^*$.

all of our experiments, our benchmark would be a direct separate Gaussianization of $\underline{X}$ and $\underline{Y}$, as an immediate alternative.

We begin with a simple toy example. Let $\underline{X} \sim N(0, I)$ and $\underline{W} \sim N(0, I)$ be independent random vectors. Define $\underline{Y} = \underline{X} + \underline{W}$, so that $\underline{X}$ and $\underline{Y}$ are jointly normally distributed. Further, we "scramble" $\underline{X}$ and $\underline{Y}$ by applying invertible, yet non-monotonic, transformations to each of them separately. We ask that the transformations are invertible to guarantee that the (analytically derived) mutual information is preserved. We further require non-monotonic transformations since marginal Gaussianization is invariant to monotonic functions (see Proposition 5), which would make this experiment too easy. In this experiment, we multiply all the observations in the range $[-1, 1]$ by -1. This operation simply mirrors these observations with respect to the origin.

Proposition 5 *Let $\tilde{X} = g(X)$ be a monotonic transformation on $X \in \mathbb{R}$. Then Gaussinizing $\tilde{X}$ is equivalent to Gaussinizing X.*

Proof Let $\tilde{V} = \Phi_N^{-1}\left(F_{\tilde{X}}\left(\tilde{X}\right)\right)$ be the Gaussianization $\tilde{X}$ and $V = \Phi_N^{-1}\left(F_X\left(X\right)\right)$ is the Gaussianization of X. Assume that g is monotonically increasing. Then,

$$F_{\tilde{X}}(a) = P(\tilde{X} \leq a) = P(g(X) \leq a) = P(X \leq g^{-1}(a)).$$

Therefore, $F_{\tilde{X}}\left(\tilde{X}\right) = F_X\left(g^{-1}(\tilde{X})\right) = F_X(X)$ and $\tilde{V} = V$. An equivalent derivation holds for the monotonically decreasing case. ■

Before we proceed, it is important to briefly comment on the implications of the finite sample size in our multivariate experiments. The ACE procedure estimates conditional expectations at each of its iterations. This estimation task is known to be quite challenging in a finite sample size regime. Breiman and Friedman (1985) suggest a *k nearest neighbor* estimator which guarantees favorable consistency properties. Unfortunately, this solution suffers from the curse of dimensionality (Hastie et al., 2005). Therefore, as the dimension of our problem increases, we cannot turn to ACE and have to settle for suboptimal solutions. In our experiments, we use the kernel CCA (Lai and Fyfe, 2000) as an alternative to ACE when the dimension size is greater than $d = 5$. The kernel CCA (KCCA) is a non-linear generalization to the classical CCA which embeds the data in a high-dimensional Hilbert space and applies CCA in that space. It is known to significantly improve the flexibility of CCA while avoiding over-fitting of the data. Notice that other non-linear CCA extensions, such as *Deep CCA* (Andrew et al., 2013) or *nonparametric CCA* (Michaeli et al., 2016), may also apply as a finite sample size alternative to ACE.

We now demonstrate our suggested approach to the jointly Gaussian model discussed above. The left plot of Figure 3 demonstrates the results we achieve for different dimension sizes d. The black line on the top is $I(\underline{X}, \underline{Y})$, which can be analytically derived. The red curve with the squares at the bottom is separate Gaussianization of $\underline{X}$ and $\underline{Y}$, which results in a very poor lower bound to the mutual information due to the non-monotonic nature of the transformation that we apply. The green curve with the circles is ACE, while the dashed blue curve is separate Gaussianization of ACE. Finally, the blue line between them is bi-terminal Gaussianization of ACE. As we can see, ACE succeeds in recovering the jointly Gaussian

representation of $\underline{X}$ and $\underline{Y}$, which makes further Gaussianization redundant. Unfortunately, for $d > 5$ we can no longer apply ACE and turn to KCCA instead. We use a Gaussian kernel with varying parameters to achieve the reported results. Since the KCCA attains a suboptimal representation it is followed by Gaussianization, which further decreases our objective. Here, we notice the improved effect of the bi-terminal Gaussianization, compared with separate Gaussianization.

Next, we turn to a more challenging exponential model. In this model, each component of $\underline{X}$ and $\underline{W}$ is exponentially distributed with a unit parameter, while all the components are independent of each other. Again, we define $\underline{Y} = \underline{X} + \underline{W}$ so that $\underline{Y}$ is Gamma distributed. This allows us to analytically derive $I(\underline{X}, \underline{Y})$. As before, we apply an invertible non-monotonic transformation to each of the components of $\underline{X}$ and $\underline{Y}$. Notice that this time we mirror the observations in the range $[0, 2]$ with respect to 1. We then apply a linear rotation, so that the components are no longer independent. The plot in the middle of Figure 3 demonstrates the results we achieve. As before, we notice that separate Gaussianzation of $\underline{X}$ and $\underline{Y}$ preforms very poorly. On the other hand, ACE as well does not succeed in maintaining the MI. This means that no Gaussianzation procedure would allow jointly normal representation of $\underline{X}$ and $\underline{Y}$ without losing information (Lemma 3). Still, by applying bi-terminal Gaussianization to ACE's results we are able to capture more than half of the information in the worst case (for $d = 5$, where ACE still applies). As before, we witness a reduction of performance when turning from ACE to KCCA.

Finally, we go back to the multivariate extension of the Gaussian mixture model described in Section 4.4 and apply our suggested procedures. Again, we witness the same behavior described in the previous experiments. In addition, our results indicate that in this model, the Gaussian part of the MI is significantly smaller, compared with the exponential model. This further demonstrates the ability of our method to quantify how well an arbitrary distribution may be represented as jointly normal.

6. Gaussian lower bound for the Information Bottleneck Curve

We now extended our derivation to the Information Bottleneck (IB) curve. We show that by maximizing the Gaussian lower bound of the mutual information (3), we allow a maximization of a Gaussian lower bound to the entire IB curve. We prove this in two steps. First, we show that the IB curve of $\phi(\underline{X}), \psi(\underline{Y})$ bounds from below the IB curve of X and Y, for any choice of ϕ, ψ (specifically, $\phi(\underline{X}) \sim N$ and $\psi(\underline{Y}) \sim N$, in our case). This property is referred to as the *data processing lemma for the IB curve*. Then, we show that the IB curve of jointly normal random variables bounds from below the IB curve of separately normal random variable. Finally, by applying the GIB (Chechik et al., 2005) to the maximally correlated jointly normal random variables that satisfy (3), we attain the desired Gaussian lower bound for the IB of $\underline{X}$ and $\underline{Y}$.

Lemma 6 *(data processing lemma for the IB Curve): Let (20) be the equivalent maximization problem of the IB problem (1):*

$$\begin{aligned} \max_{T} \quad & I(T(\underline{X}); \underline{Y}) \\ \textit{subject to} \quad & I(T(\underline{X}); \underline{X}) \leq I_X. \end{aligned} \tag{20}$$

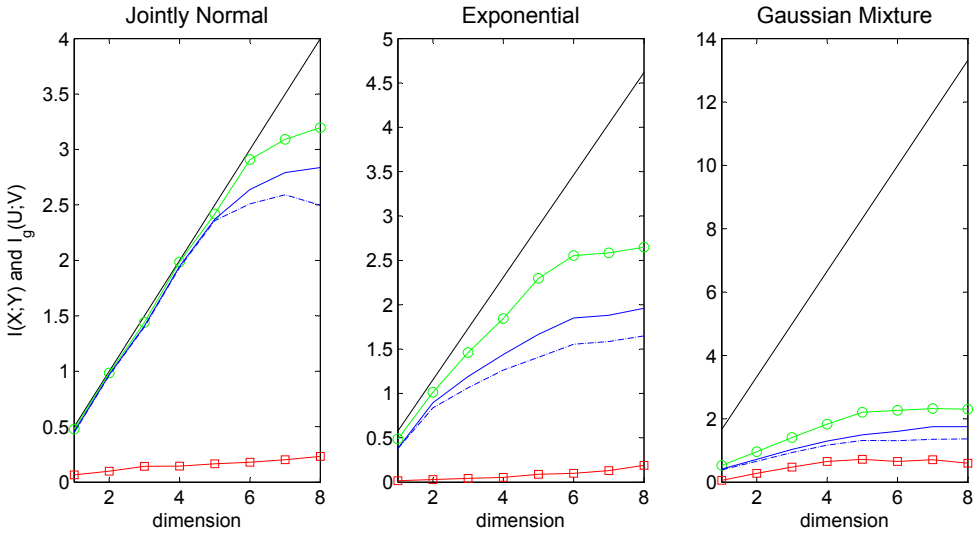

Figure 3: Multivariate Gaussianization experiments: The black line on the top of each plot is $I(\underline{X}, \underline{Y})$. The red curve with the squares at the bottom is separate Gaussianization of $\underline{X}$ and $\underline{Y}$. The green curve with the circles is ACE, while the dashed blue curve is separate Gaussianization of ACE. The blue line in between is bi-terminal Gaussianization of ACE.

Denote its solution as $I_*^\beta(\underline{X};\underline{Y})$. *Then,* $I_*^\beta(\underline{X};\underline{Y}) \geq I_*^\beta(\phi(\underline{X});\psi(\underline{Y}))$ *for any* ϕ, ψ, *and with equality iff* $I(\underline{X};\underline{Y}) = I(\phi(\underline{X});\psi(\underline{Y}))$.

Proof We prove this lemma by showing that $I_*^\beta(\underline{X};\underline{Y}) \geq I_*^\beta(\underline{X};\psi(\underline{Y})) \geq I_*^\beta(\psi(\underline{X});\psi(\underline{Y}))$. We start with the first inequality. According to the data processing lemma, we have that $I(T(\underline{X});\underline{Y}) \geq I(T(\underline{X});\psi(\underline{Y}))$. Notice that for convenience, we emphasize that T is indeed a mapping of X alone. In addition, since our constraint (1) is independent of Y, we have that $I_*^\beta(\underline{X};\underline{Y}) \geq I_*^\beta(\underline{X};\psi(\underline{Y}))$, as desired. Second, denote the minimizer of (1) as $I_*(\underline{X};\underline{Y})$. Assume that there exists such ϕ that

$$I_*(\underline{X};\underline{Y}) > I_*(\phi(\underline{X});\underline{Y}). \tag{21}$$

This means that for $I(T(\underline{X});\underline{Y}) \geq I_Y$ and $I(T'(\phi(\underline{X}));\underline{Y}) \geq I_Y$ we have that $I(T(\underline{X});\underline{X}) > I(T'(\phi(\underline{X}));\phi(\underline{X}))$ where T and T' are the optimizers of (1) with respect to $(\underline{X},\underline{Y})$ and $(\phi(\underline{X}),\underline{Y})$, for a given I_Y, respectively. Let us set $\bar{T} \equiv T' \circ \phi$ and apply this transformation to $\underline{X}$. Then, we have that the constraint of (1) is met, as $I(\bar{T}(\underline{X});\underline{Y}) \equiv I(T'(\phi(\underline{X}));\underline{Y}) \geq I_Y$. In addition, we have that

$$I(\bar{T}(\underline{X});\underline{X}) \equiv I(T'(\phi(\underline{X}));\underline{X}) = I(T'(\phi(\underline{X}));\phi(\underline{X}))$$

where the second equality follows from T' being independent of $\underline{X}$, given $\phi(\underline{X})$. Therefore, $\bar{T} = T' \circ \phi$ is a better optimizer to (1) with respect to $\underline{X}$ and $\underline{Y}$, then T. This contradicts the optimality of T as a minimizer of (1), which means that the assumption in (21) is false. Therefore, $I_*(\underline{X};\underline{Y}) \leq I_*(\phi(\underline{X});\underline{Y})$ which means that $I_*^\beta(\underline{X};\underline{Y}) \geq I_*^\beta(\phi(\underline{X});\underline{Y})$ for any $\underline{Y}$ (specifically, $\phi(\underline{Y})$) and our proof is concluded. ■

Lemma 7 *Let $\underline{U}$ and $\underline{V}$ be separately Gaussian random vectors with a joint covariance matrix $C_{[\underline{U},\underline{V}]}$ (that is, $\underline{U} \sim N$ and $\underline{V} \sim N$ but $[\underline{U},\underline{V}]^T$ is not normally distributed). Let $\underline{U}_{jg}, \underline{V}_{jg}$ be two jointly normally distributed random vectors with the same covariance matrix, $C_{[\underline{U}_{jg},\underline{V}_{jg}]} = C_{[\underline{U},\underline{V}]}$. Then, the IB curve of $\underline{U}_{jg}$ and $\underline{V}_{jg}$ bounds from below the IB curve of $\underline{U}$ and $\underline{V}$.*

Proof Let $\big(I(\underline{U}_{jg};\underline{T}), I(\underline{T};\underline{V}_{jg})\big)$ be a point of the IB curve of $\underline{U}_{jg}$ and $\underline{V}_{jg}$. Since $\underline{U}_{jg}$ and $\underline{V}_{jg}$ are jointly normally distributed, T is necessarily a linear transformation of $\underline{U}_{jg}$, with additive independent Gaussian noise (Chechik et al., 2005). Specifically, $T = A\underline{U}_{jg} + \underline{\zeta}$, where $\underline{\zeta} \sim N(0, I)$, independent of $\underline{U}_{jg}$ and $\underline{V}_{jg}$.

Further, let $\underline{T}' = A\underline{U} + \underline{\zeta}$ be the same transformation, applied of $\underline{U}$. Since $\underline{U}$ and $\underline{V}$ are not jointly normal, the point $(I(\underline{U};\underline{T}'), I(\underline{T}';\underline{V}))$ is below the IB curve of $\underline{U}$ and $\underline{V}$. First, notice that

$$I(\underline{U};\underline{T}') \equiv I(\underline{U}; A\underline{U} + \underline{\zeta}) = I(\underline{U}_{jg}; A\underline{U}_{jg} + \underline{\zeta}) \equiv I(\underline{U}_{jg}, \underline{T})$$

where the second equality follows from $\underline{U}$ and $\underline{U}_{jg}$ having the same distribution. In addition, since $C_{[\underline{U}_{jg},\underline{V}_{jg}]} = C_{[\underline{U},\underline{V}]}$ we have that $C_{[A\underline{U}_{jg}+\underline{\zeta},\underline{V}_{jg}]} = C_{[A\underline{U}+\underline{\zeta},\underline{V}]}$. Therefore, $I(A\underline{U} + \underline{\zeta};\underline{V}) \geq I(A\underline{U}_{jg} + \underline{\zeta};\underline{V}_{jg})$, in the same manner as the in (3). This means that $I(\underline{T}';\underline{V}) \geq I(\underline{T};\underline{V}_{jg})$. To conclude, we showed that for the two pairs, $\big(I(\underline{U}_{jg};\underline{T}), I(\underline{T};\underline{V}_{jg})\big)$ and $(I(\underline{U};\underline{T}'), I(\underline{T}';\underline{V}))$, we have that $I(\underline{U};\underline{T}') = I(\underline{U}_{jg},\underline{T})$ while $I(\underline{T}';\underline{V}) \geq I(\underline{T};\underline{V}_{jg})$, as desired. ■

The two theorems above guarantee that the IB curve of $\underline{X}$ and $\underline{Y}$ is bounded from below by the IB curve of $\underline{U}_{jg}$ and $\underline{V}_{jg}$, where $C_{[\underline{U}_{jg},\underline{V}_{jg}]} = C_{[\underline{U},\underline{V}]}$, and $\underline{U} = \phi(\underline{X}) \sim N$, $\underline{V} = \psi(\underline{Y}) \sim N$. Therefore, in order to maximize this lower bound, one needs to maximize the correlation between $\underline{U}$ and $\underline{V}$, subject to a normality constraint, as discussed throughout this manuscript. Moreover, once we have found a pair of $(\underline{U}_{jg}, \underline{V}_{jg})$ with a maximal correlation, we may directly apply the GIB to it, as shown by Chechik et al. (2005), to achieve the optimal Gaussian lower bound oIB curve for $\underline{X}$ and $\underline{Y}$.

6.1 Examples

We now demonstrate our suggested Gaussian lower bound for the IB curve in two different setups. Here, we would like to compare our bound with the "true" IB curve, and with an additional benchmark off-the-shelf lower bound. As discussed in Section 1, computing the exact IB curve (for a general joint distribution) is not a simple task. This task becomes even more complicated when dealing with continuous random variables. In fact, to the best of our knowledge, all currently known methods provide approximated curves, which do not claim to converge to the exact IB curve. Moreover, these methods fail to provide any guarantees on the extent of their divergence from the true IB curve. Therefore, in our experiments, we apply the commonly used reverse annealing technique (Slonim, 2002) in order to approximate the "true" IB curve. The reverse annealing algorithm is initiated by computing the mutual information between $\underline{X}$ and $\underline{Y}$, which corresponds to extreme point where $I_Y \to \infty$ on the IB curve. Then, I_Y is gradually decreased and the solution of the IB problem (1) with the previous value of I_Y serves as a starting point to the currently solved I_Y. This results in a greedy "no-regret" optimization method, which in general, fails to

converge to the exact IB curve. However, in some special cases (such as the GIB), it can be shown that the optimal solution for a given value of I_Y is, in fact, the optimal starting point for a smaller value of I_Y. In the general case, it is implicitly assumed to be a reasonable local optimization domain. Since the reverse annealing was originally designed for discrete random variables, we apply discretization (via Gaussian quadratures) to our probability distributions is all of our experiments.

We begin by revisiting the exponential model, described in Section 5.4. In this model, X and W are independent exponentially distributed random variables with a unit parameter. We define $Y = X + W$ so that Y is Gamma distributed. As in Section 5.4 we apply an invertible non-monotonic transformation to X and Y, to make this problem more challenging. Since approximating the IB curve is involved enough for continuous random variables, we limit our attention to the simplest univariate case.

The plot on the left of Figure 4 demonstrates the results we achieve. The black curve on top is the approximated IB curve, using the reverse annealing procedure. The red curve on the bottom is a benchmark lower bound, achieved by simply applying the GIB to X and Y, as if they were jointly Gaussian. The blue curve in the middle is our suggested Gaussian lower bound (Section 4.2). As we can see, our suggested bound surpasses the GIB quite remarkably. This is mainly due to the non-monotonic transformation we apply, which makes the joint distribution highly non-Gaussian. We further notice that our bound is quite tight for smaller I_Y's (closer to the origin) but increasingly diverges as I_Y increases. The reason is that more compressed representations are more "degenerate" and are easier to Gaussianize while maintaining reasonably high correlations.

Next, we revisit the more challenging Gaussian mixture model, described in Section 4.4. The right plot in Figure 4 demonstrates the results we achieve. As before, we notice that our suggested lower bound surpasses the naive benchmark, while demonstrating favorable performance closer to the origin. Comparing the two models, we notice that the Gaussian mixture is more difficult to bound from below using our suggested method. This result is not surprising, given the gap in our ability to bound from below the mutual information in these two models, as discussed in Section 5.4.

7. Discussion and conclusion

In this work we address the fundamental problem of normalizing non-Gaussian data, while trying to avoid loss of information. This allows us to solve complex problems by linear means, as we push information to the data's second moments. We show that our ability to do so is strongly governed by the non-linear canonical correlations of the data. In other words, if the non-linear canonical coefficients of the data fail to maintain its mutual information, then it is impossible to describe its high order dependencies just by second order statistics. This result is of high interest to a broad variety of applications, as solving non-linear problems by linear means is a common alternative in many scientific and engineering fields. Further, we provide a variety of methods to quantify the minimal amount of information that may be lost when normalizing the data. We show that in many cases, our suggested approach is able to preserve a significant portion of the information, even for highly non-Gaussian joint distributions. Our results improve upon Cardoso (2003) information geometry bound, as we show that a tighter bound may be obtained by the AGCE method.

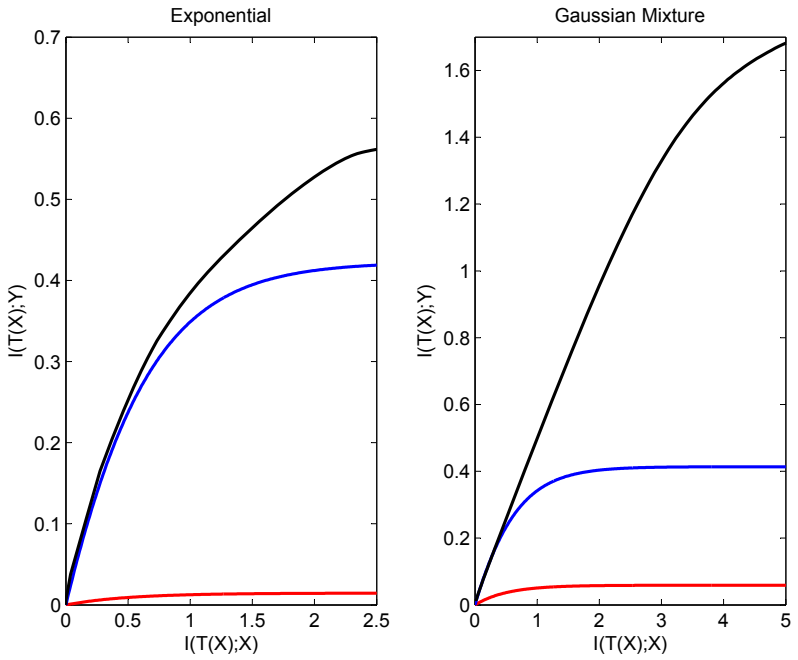

Figure 4: Bounding the Information Bottleneck curve for Exponential and Gaussian Mixture distributions: The black line is the approximated IB curve and the blue line is our suggested Gaussian lower bound. The red curve is achieved by applying the GIB directly to X, Y.

It is important to mention that while our suggested approach is theoretically found, it exhibits several practical limitation in a finite sample-size setup. This is a direct result of our use of the ACE algorithm, which suffers from the curse of dimensionality when applied to high-dimensional data. Therefore, we further examine different non-linear CCA methods, which are less vulnerable to this problem. However, these methods fail to converge to the optimal canonical coefficients.

Finally, we show that our results may be generalized to bound from below the entire information bottleneck curve. This allows a practical alternative for different approximation methods and restrictive solutions to the involved IB problem in the continuous case. Our experiments show that the suggested Gaussian lower bound provides a meaningful benchmark to the IB curve, even in highly non-Gaussian setups.

8. Acknowledgments

This research was supported by a Fellowship from the Israeli Center of Research Excellence in Algorithms to Amichai Painsky. The authors thank Nori Jacoby for early discussions on the subject.

References

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine*

Intelligence, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

Leo Breiman and Jerome H Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985.

Jean-François Cardoso. Dependence, correlation and gaussianity in independent component analysis. *Journal of Machine Learning Research*, 4:1177–1203, 2003.

Matthew Chalk, Olivier Marre, and Gasper Tkacik. Relevant sparse codes with variational information bottleneck. In *Advances in Neural Information Processing Systems*, pages 1957–1965, 2016.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165–188, 2005.

Scott Saobing Chen and Ramesh A Gopinath. Gaussianization. In *Advances in Neural Information Processing Systems*, pages 423–429, 2001.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Nir Friedman, Ori Mosenzon, Noam Slonim, and Naftali Tishby. Multivariate information bottleneck. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 152–161. Morgan Kaufmann Publishers Inc., 2001.

Walter R Gilks. *Markov chain monte carlo*. Wiley Online Library, 2005.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

Ron M Hecht, Elad Noor, and Naftali Tishby. Speaker recognition by Gaussian information bottleneck. In *INTERSPEECH*, pages 1567–1570, 2009.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

Jim Kay. Feature discovery under contextual supervision using mutual information. In *Neural Networks, 1992. IJCNN., International Joint Conference on*, volume 4, pages 79–84. IEEE, 1992.

Arto Klami and Samuel Kaski. Non-parametric dependent components. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, 2005.

Artemy Kolchinsky, Brendan D Tracey, and David H Wolpert. Nonlinear information bottleneck. *arXiv preprint arXiv:1705.02436*, 2017.

Pei Ling Lai and Colin Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(05):365–377, 2000.

HO Lancaster. Correlations and canonical forms of bivariate distributions. *The Annals of Mathematical Statistics*, 34(2):532–538, 1963.

Valero Laparra, Gustavo Camps-Valls, and Jesús Malo. Iterative Gaussianization: from ICA to random rotations. *IEEE Transactions on Neural Networks*, 22(4):537–549, 2011.

Tomer Michaeli, Weiran Wang, and Karen Livescu. Nonparametric canonical correlation analysis. In *International Conference on Machine Learning*, pages 1967–1976, 2016.

Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media, 1998.

Mélanie Rey and Volker Roth. Meta-Gaussian information bottleneck. In *Advances in Neural Information Processing Systems*, pages 1916–1924, 2012.

Elad Schneidman, Noam Slonim, Naftali Tishby, R deRuyter van Steveninck, and William Bialek. Analyzing neural codes using the information bottleneck method. *Advances in Neural Information Processing Systems*, 2001.

Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440 (7087):1007–1012, 2006.

Ofer Shayevitz and Meir Feder. Optimal feedback communication via posterior matching. *IEEE Transactions on Information Theory*, 57(3):1186–1222, 2011.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Janne Sinkkonen and Samuel Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14(1):217–239, 2002.

Noam Slonim. *The information bottleneck: Theory and applications*. PhD thesis, Hebrew University of Jerusalem, 2002.

Noam Slonim and Naftali Tishby. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 208–215. ACM, 2000.

Noam Slonim, Gurinder Singh Atwal, Gašper Tkačik, and William Bialek. Information-based clustering. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18297–18302, 2005.

Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer, 2011.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proceedings of 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.