1

Information Bottleneck for Gaussian Variables

Gal Chechik*† gal@www.stanford.edu
http://robotics.stanford.edu/∼gal/

Amir Globerson* gamir@cs.huji.ac.il

Naftali Tishby tishby@cs.huji.ac.il

Yair Weiss yweiss@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: Peter Dayan

Abstract

The problem of extracting the relevant aspects of data was previously addressed through the *information bottleneck* (IB) method, through (soft) clustering one variable while preserving information about another - *relevance* - variable. The current work extends these ideas to obtain continuous representations that preserve relevant information, rather than discrete clusters, for the special case of multivariate Gaussian variables. While the general continuous IB problem is difficult to solve, we provide an analytic solution for the optimal representation and tradeoff between compression and relevance for the this important case. The obtained optimal representation is a noisy linear projection to eigenvectors of the normalized regression matrix $\Sigma_{x|y}\Sigma_x^{-1}$, which is also the basis obtained in Canonical Correlation Analysis. However, in Gaussian IB, the compression tradeoff parameter uniquely determines the dimension, as well as the scale of each eigenvector, through a cascade of structural phase transitions. This introduces a novel interpretation where solutions of different ranks lie on a continuum parametrized by the compression level. Our analysis also provides a complete analytic expression of the preserved information as a function of the compression (the "information-curve"), in terms of the eigenvalue spectrum of the data. As in the discrete case, the information curve is concave and smooth, though it is made of different analytic segments for each optimal dimension. Finally, we show how the algorithmic theory developed in the IB framework provides an iterative algorithm for obtaining the optimal Gaussian projections.

Keywords: Information Bottleneck, Gaussian Processes, Dimensionality Reduction, Canonical Correlation Analysis

∗. Both authors contributed equally.
† Corresponding author. Current address: CS Dept., Stanford University, Stanford CA 94305-9025

1. Introduction

Extracting relevant aspects of complex data is a fundamental task in machine learning and statistics. The problem is often that the data contains many structures, which make it difficult to define which of them are relevant and which are not in an unsupervised manner. For example, speech signals may be characterized by their volume level, pitch, or content; pictures can be ranked by their luminosity level, color saturation or importance with regard to some task.

This problem was addressed in a principled manner by the information bottleneck (IB) approach (Tishby et al., 1999). Given the joint distribution of a "source" variable X and another "relevance" variable Y, IB operates to compress X, while preserving information about Y. The variable Y thus implicitly defines what is relevant in X and what is not. Formally, this is cast as the following variational problem

$$\min_{p(t|x)} \mathcal{L} : \mathcal{L} \equiv I(X;T) - \beta I(T;Y) \tag{1}$$

where T represents the compressed representation of X via the conditional distributions $p(t|x)$, while the information that T maintains on Y is captured by the distribution $p(y|t)$. This formulation is general and does not depend on the type of the X, Y distribution. The positive parameter β determines the tradeoff between compression and preserved relevant information, as the Lagrange multiplier for the constrained optimization problem $\min_{p(t|x)} I(X;T) - \beta\left(I(T;Y) - const\right)$. Since T is a function of X it is independent of Y given X, thus the three variables can be written as the Markov chain $Y - X - T$. From the information inequality we thus have $I(X;T) - \beta I(T;Y) \geq (1-\beta) I(T;Y)$, and therefore for all values of $\beta \leq 1$, the optimal solution of the minimization problem is degenerated $I(T;X) = I(T;Y) = 0$. As we will show below, the range of degenerated solutions is even larger for Gaussian variables and depends on the eigen spectrum of the variables covariance matrices.

The rationale behind the IB principle can be viewed as model-free "looking inside the black-box" system analysis approach. Given the input-output (X, Y) "black-box" statistics, IB aims to construct efficient representations of X, denoted by the variable T, that can account for the observed statistics of Y. IB achieves this using a single tradeoff parameter to represent the tradeoff between the complexity of the representation of X, measured by $I(X;T)$, and the accuracy of this representation, measured by $I(T;Y)$. The choice of mutual information for the characterization of complexity and accuracy stems from Shannon's theory, where information minimization corresponds to optimal compression in Rate Distortion Theory, and its maximization corresponds to optimal information transmission in Noisy Channel Coding.

From a machine learning perspective, IB may be interpreted as regularized generative modeling. Under certain conditions $I(T;Y)$ can be interpreted as an empirical likelihood of a special mixture model, and $I(T;X)$ as penalizing complex models (Slonim and Weiss, 2002). While this interpretation can lead to interesting analogies , it is important to emphasize the differences. First, IB views $I(X;T)$ not as a regularization term, but rather corresponds to the distortion constraint in the original system. As a result, this constraint is useful even when the joint distribution is known exactly, because the goal of IB is to obtain compact representations rather than to estimate density. Interestingly, $I(T;X)$ also characterizes the complexity of the representation T as the expected number of bits needed to specify the t for a given x. In that role it can be viewed as an expected "cost" of the internal representation, as in MDL. As is well acknowledged now source coding with distortion and channel coding with cost are dual problems (see for example Shannon, 1959, Pradhan et al., 2003). In that information theoretic sense, IB is *self dual*, where the resulting source and channel are perfectly matched (as in Gastpar and Vetterli, 2003).

The information bottleneck approach has been applied so far mainly to categorical variables, with a discrete T that represents (soft) clusters of X. It has been proved useful for a range of applications from documents clustering (Slonim and Tishby, 2000) through neural code analysis (Dimitrov and Miller, 2001) to gene expression analysis (Friedman et al., 2001, Sinkkonen and Kaski, 2001) (for a more detailed review of IB clustering algorithms see Slonim (2003)). However, its general information theoretic formulation is not restricted, both in terms of the nature of the variables X and Y, as well as of the compression variable T. It can be naturally extended to nominal, categorical, and continuous variables, as well as to dimension reduction rather than clustering techniques. The goal of this paper is apply the IB for the special, but very important, case of Gaussian processes which has become one of the most important generative classes in machine learning. In addition, this is the first concrete application of IB to dimension reduction with continuous compressed representation, and as such exhibit interesting dimension related phase transitions.

The general solution of IB for continuous T yields the same set of self-consistent equations obtained already in (Tishby et al., 1999), but solving these equations for the distributions $p(t|x)$, $p(t)$ and $p(y|t)$ without any further assumptions is a difficult challenge, as it yields non-linear coupled eigenvalue problems. As in many other cases, however, we show here that the problem turns out to be analytically tractable when X and Y are joint multivariate Gaussian variables. In this case, rather than using the fixed point equations and the generalized Blahut-Arimoto algorithm as proposed in (Tishby et al., 1999), one can explicitly optimize the target function with respect to the mapping $p(t|x)$ and obtain a closed form solution of the optimal dimensionality reduction.

The optimal compression in the Gaussian Information Bottleneck (GIB) is defined in terms of the compression-relevance tradeoff (also known as the "Information Curve", or "Accuracy-Complexity" tradeoff), determined by varying the parameter β. The optimal solution turns out to be a noisy linear projection to a subspace whose dimensionality is determined by the parameter β. The subspaces are spanned by the basis vectors obtained as in the well known *Canonical Correlation Analysis* (CCA) (Hotelling, 1935), but the exact nature of the projection is determined in a unique way via the parameter β. Specifically, as β increases, additional dimensions are added to the projection variable T, through a series of critical points (structural phase transitions), while at the same time the relative magnitude of each basis vector is rescaled. This process continues until all the relevant information about Y is captured in T. This demonstrates how the IB principle can provide a continuous measure of model complexity in information theoretic terms.

The idea of maximization of relevant information was also taken in the *Imax* framework of Becker and Hinton (Becker and Hinton, 1992, Becker, 1996), which followed Linsker's idea of information maximization (Linsker, 1988, 1992). In the Imax setting, there are two one-layer feed forward networks with inputs X_a, X_b and outputs neurons Y_a, Y_b; the output neuron Y_a serves to define relevance to the output of the neighboring network Y_b.

Formally, the goal is to tune the incoming weights of the output neurons, such that their mutual information $I(Y_a; Y_b)$ is maximized. An important difference between *Imax* and the IB setting, is that in the *Imax* setting, $I(Y_a; Y_b)$ is invariant to scaling and translation of the Y's since the compression achieved in the mapping $X_a \rightarrow Y_a$ is not modeled explicitly. In contrast, the IB framework aims to characterize the dependence of the solution on the explicit compression term $I(T; X)$, which is a *scale sensitive* measure when the transformation is noisy. This view of compressed representation T of the inputs X is useful when dealing with neural systems that are stochastic in nature and limited in their responses amplitudes and are thus constrained to finite $I(T; X)$.

The current paper starts by defining the problem of relevant information extraction for Gaussian variables. Section 3 gives the main result of the paper: an analytical characterization of the optimal projections, which is then developed in Section 4. Section 5 develops an analytical expression for the GIB compression-relevance tradeoff - the information curve. Section 6 shows how the general IB algorithm can be adapted to the Gaussian case, yielding an iterative algorithm for finding the optimal projections. The relations to canonical correlation analysis and coding with side-information are discussed in Section 9.

2. Gaussian Information Bottleneck

We now formalize the problem of Information Bottleneck for Gaussian variables. Let (X, Y) be two jointly multivariate Gaussian variables of dimensions n_x, n_y and denote by Σ_x,Σ_y the covariance matrices of X,Y and by Σ_{xy} their cross-covariance matrix [1]. The goal of GIB is to compress the variable X via a stochastic transformation into another variable $T \in R^{n_x}$, while preserving information about Y. The dimension of T is not explicitly limited in our formalism, since we will show that the effective dimension is determined by the value of β.

It is shown in Globerson and Tishby (2004) that the optimum for this problem is obtained by a variable T which is also jointly Gaussian with X. The formal proof uses the entropy power inequality as in Berger and Zamir (1999), and is rather technical, but an intuitive explanation is that since X and Y are Gaussians, the only statistical dependencies that connect them are bi-linear. Therefore, a linear projection of X is sufficient to capture all the information that X has on Y. The Entropy-power inequality is used to show that a linear projection of X, which is also Gaussian in this case, indeed attains this maximum information.

Since every two centered random variables X and T with jointly Gaussian distribution can be presented through the linear transformation $T = AX + \xi$, where $\xi \sim N(0, \Sigma_\xi)$ is another Gaussian that is independent of X, we formalize the problem using this representation of T, as the following minimization,

$$\min_{A,\Sigma_\xi} \mathcal{L} \equiv I(X; T) - \beta I(T; Y) \tag{2}$$

over the noisy linear transformations of A, Σ_ξ

$$T = AX + \xi; \quad \xi \sim N(0, \Sigma_\xi) \ . \tag{3}$$

1. For simplicity we assume that X and Y have zero means and Σ_x, Σ_y are full rank. Otherwise X and Y can be centered and reduced to the proper dimensionality.

Thus T is normally distributed $T \sim N(0, \Sigma_t)$ with $\Sigma_t = A\Sigma_x A^T + \Sigma_\xi$.

Interestingly, the term ξ can also be viewed as an additive noise term, as commonly done in models of learning in neural networks. Under this view, ξ serves as a regularization term whose covariance determines the scales of the problem. While the goal of GIB is to find the optimal projection parameters A, Σ_ξ jointly, we show below that the problem factorizes such that the optimal projection A does not depend on the noise, which does not carry any information about Y.

3. The Optimal Projection

The first main result of this paper is the characterization of the optimal A,Σ_ξ as a function of β

Theorem 3.1 *The optimal projection $T = AX + \xi$ for a given tradeoff parameter β is given by $\Sigma_\xi = I_x$ and*

$$A = \left\{ \begin{array}{cc} \left[\mathbf{0}^T; \ldots; \mathbf{0}^T\right] & 0 \le \beta \le {\beta^c}_1 \\ \left[\alpha_1 \mathbf{v}_1^T, \mathbf{0}^T; \ldots; \mathbf{0}^T\right] & \beta_1^c \le \beta \le {\beta^c}_2 \\ \left[\alpha_1 \mathbf{v}_1^T; \alpha_2 \mathbf{v}_2^T; \mathbf{0}^T; \ldots; \mathbf{0}^T\right] & {\beta^c}_2 \le \beta \le {\beta^c}_3 \\ \vdots & \end{array} \right\} \tag{4}$$

where $\{\mathbf{v}_1^T, \mathbf{v}_2^T, \ldots, \mathbf{v}_{n_x}^T\}$ are left eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$ sorted by their corresponding ascending eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{n_x}$, ${\beta^c}_i = \frac{1}{1-\lambda_i}$ are critical β values, α_i are coefficients defined by $\alpha_i \equiv \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$, $r_i \equiv \mathbf{v}_i^T \Sigma_x \mathbf{v}_i$, $\mathbf{0}^T$ is an n_x dimensional row vector of zeros, and semicolons separate rows in the matrix A.

This theorem asserts that the optimal projection consists of eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$, combined in an interesting manner: For β values that are smaller than the smallest critical point ${\beta^c}_1$, compression is more important than any information preservation and the optimal solution is the degenerated one $A \equiv 0$. As β is increased, it goes through a series of critical points ${\beta^c}_i$, at each of which another eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ is added to A. Even though the rank of A increases at each of these transition points, A changes continuously as a function of β since at each critical point ${\beta^c}_i$ the coefficient α_i vanishes. Thus β parameterizes a sort of “continuous rank” of the projection.

To illustrate the form of the solution, we plot the landscape of the target function $\mathcal{L}$ together with the solution in a simple problem where $X \in R^2$ and $Y \in R$. In this case A has a single non-zero row, thus A can be thought of as a row vector of length 2, that projects X to a scalar $A : X \to R$, $T \in R$. Figure 1 shows the target function $\mathcal{L}$ as a function of the (vector of length 2) projection A. In this example, the largest eigenvalue is $\lambda_1 = 0.95$, yielding ${\beta^c}_1 = 20$. Therefore, for $\beta = 15$ (Figure 1A) the zero solution is optimal, but for $\beta = 100 > \beta^c$ (Figure 1B) the corresponding eigenvector is a feasible solution, and the target function manifold contains two mirror minima. As β increases from 1 to ∞, these two minima, starting as a single unified minimum at zero, split at β^c, and then diverge apart to ∞.

We now turn to prove theorem 3.1.

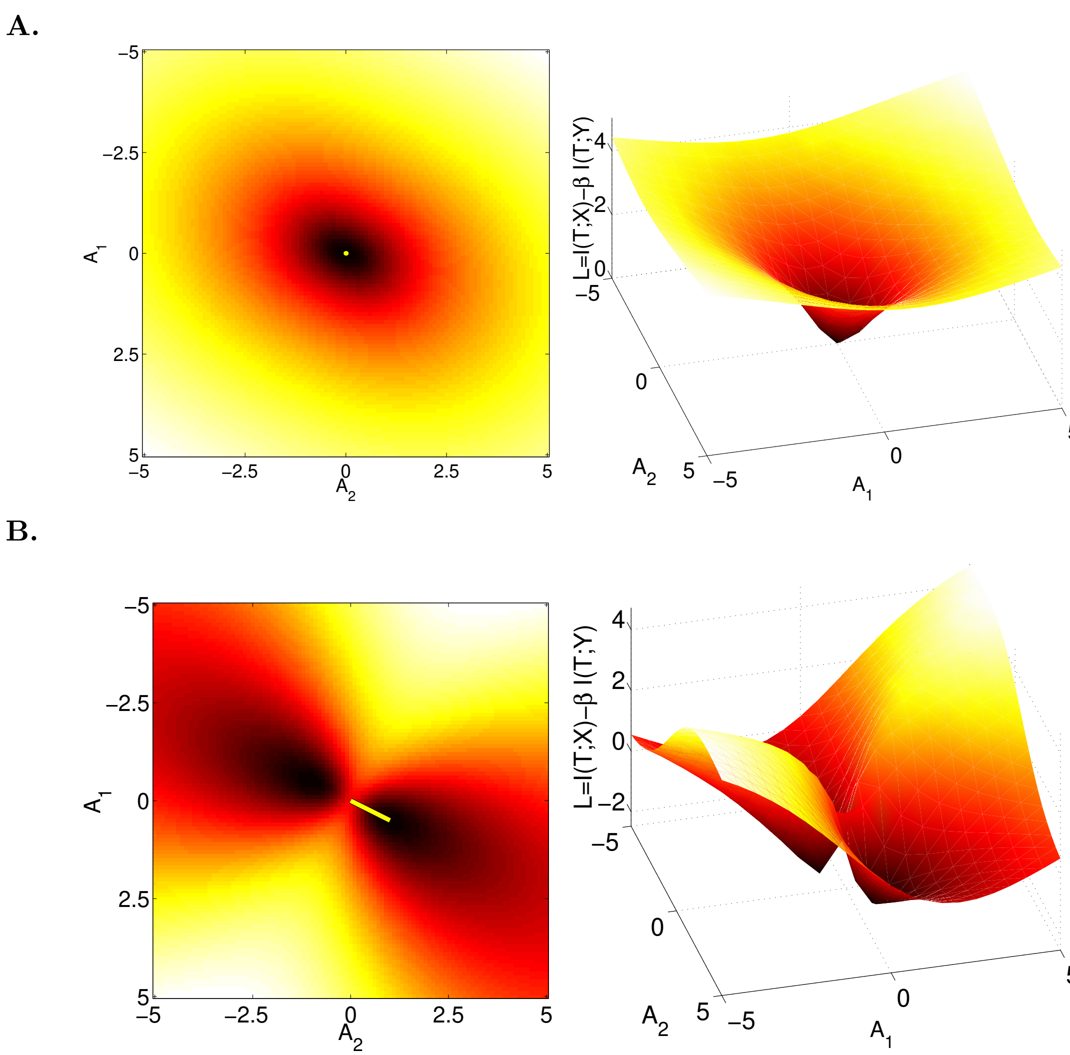

Figure 1: The surface of the target function $\mathcal{L}$ calculated numerically as a function of the optimization parameters in two illustrative examples with a scalar projection $A : R^2 \rightarrow R$. Each row plots the target surface $\mathcal{L}$ both in 2D (left) and 3D (right) as a function of the (two dimensional) projections A. **A.** For $\beta = 15$, the optimal solution is the degenerated solution $A \equiv 0$. **B.** For $\beta = 100$, a non degenerate solution is optimal, together with its mirror solution. The $\Sigma_{x|y}\Sigma_x^{-1}$- eigenvector of smallest eigenvalue, with a norm computed according to theorem 3.1 is superimposed, showing that it obtains the global minimum of $\mathcal{L}$. Parameters' values $\Sigma_{xy} = [0.1\ 0.2]$, $\Sigma_x = I_2$, $\Sigma_\xi = 0.3 I_{2\times 2}$.

4. Deriving the Optimal Projection

We first rewrite the target function as

$$\mathcal{L} = I(X;T) - \beta I(T;Y) = h(T) - h(T|X) - \beta h(T) + \beta h(T|Y) \tag{5}$$

where h is the (differential) entropy of a continuous variable

$$h(X) \equiv -\int_X f(x) \log f(x)\, dx \quad .$$

Recall that the entropy of a d dimensional Gaussian variable is

$$h(X) = \frac{1}{2} \log\left((2\pi e)^d |\Sigma_x|\right)$$

where $|x|$ denotes the determinant of x, and Σ_x is the covariance of X. We therefore turn to calculate the relevant covariance matrices. From the definition of T we have $\Sigma_{tx} = A\Sigma_x$, $\Sigma_{ty} = A\Sigma_{xy}$ and $\Sigma_t = A\Sigma_x A^T + \Sigma_\xi$. Now, the conditional covariance matrix $\Sigma_{x|y}$ can be used to calculate the covariance of the conditional variable $T|Y$, using the Schur complement formula (see e.g., Magnus and Neudecker, 1988)

$$\Sigma_{t|y} = \Sigma_t - \Sigma_{ty}\Sigma_y^{-1}\Sigma_{yt} = A\Sigma_{x|y}A^T + \Sigma_\xi$$

The target function can now be rewritten as

$$\begin{aligned} \mathcal{L} &= \log(|\Sigma_t|) - \log(|\Sigma_{t|x}|) - \beta \log(|\Sigma_t|) + \beta \log(|\Sigma_{t|y}|) \\ &= (1-\beta)\log(|A\Sigma_x A^T + \Sigma_\xi|) - \log(|\Sigma_\xi|) + \beta \log(|A\Sigma_{x|y}A^T + \Sigma_\xi|) \end{aligned} \tag{6}$$

Although $\mathcal{L}$ is a function of both the noise Σ_ξ *and* the projection A, Lemma A.1 in Appendix A shows that for every pair (A, Σ_ξ), there is another projection $\tilde{A}$ such that the pair $(\tilde{A}, I)$ obtains the same value of $\mathcal{L}$. This is obtained by setting $\tilde{A} = \sqrt{D^{-1}}VA$ where $\Sigma_\xi = VDV^T$, which yields $\mathcal{L}(\tilde{A}, I) = \mathcal{L}(A, \Sigma_\xi)$[2]. This allows us to simplify the calculations by replacing the noise covariance matrix Σ_ξ with the identity matrix I_d.

To identify the minimum of $\mathcal{L}$ we differentiate $\mathcal{L}$ w.r.t. to the projection A using the algebraic identity $\frac{\delta}{\delta A}\log(|ACA^T|) = (ACA^T)^{-1}2AC$ which holds for any symmetric matrix C.

$$\frac{\delta \mathcal{L}}{\delta A} = (1-\beta)(A\Sigma_x A^T + I_d)^{-1} 2A\Sigma_x + \beta(A\Sigma_{x|y}A^T + I_d)^{-1}2A\Sigma_{x|y} \tag{7}$$

Equating this derivative to zero and rearranging, we obtain necessary conditions for an internal minimum of $\mathcal{L}$, which we explore in the next two sections.

4.1 Scalar projections

For clearer presentation of the general derivation, we begin with a sketch of the proof by focusing on the case where T is a scalar, that is, the optimal projection matrix A is a now a single row vector. In this case, both $A\Sigma_x A^T$ and $A\Sigma_{x|y}A^T$ are scalars, and we can write

$$\left(\frac{\beta - 1}{\beta}\right)\left(\frac{A\Sigma_{x|y}A^T + 1}{A\Sigma_x A^T + 1}\right) A = A\left[\Sigma_{x|y}\Sigma_x^{-1}\right] \quad . \tag{8}$$

2. Although this holds only for full rank Σ_ξ, it does not limit the generality of the discussion since low rank matrices yield infinite values of $\mathcal{L}$ and are therefore suboptimal.

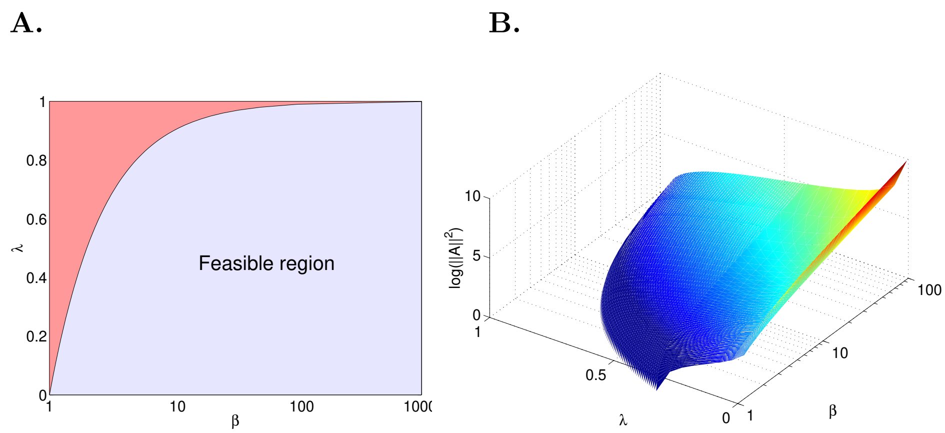

Figure 2: **A.** The regions of (β,λ) pairs that lead to the zero (red) and eigenvector (blue) solutions. **B.** The norm $||A||^2$ as a function of β and λ over the feasible region.

This equation is therefore an eigenvalue problem in which the eigenvalues depend on A. It has two types of solutions depending on the value of β. First, A may be identically zero. Otherwise, A must be the eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$, with an eigenvalue $\lambda = \frac{\beta-1}{\beta}\frac{A\Sigma_{x|y}A^T+1}{A\Sigma_x A^T+1}$

To characterize the values of β for which the optimal solution does not degenerate, we find when the eigenvector solution is optimal. Denote the norm of Σ_x w.r.t. A by $r = \frac{A\Sigma_x A^T}{||A||^2}$. When A is an eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$, Lemma B.1 shows that r is positive and that $A\Sigma_{x|y}\Sigma_x^{-1}\Sigma_x A^T = \lambda r||A||^2$. Rewriting the eigenvalue and isolating $||A||^2$, we have

$$0 < ||A||^2 = \frac{\beta(1-\lambda)-1}{r\lambda} \ . \tag{9}$$

This inequality provides a constraint on β and λ that is required for a non-degenerated type of solution

$$\lambda \leq \frac{\beta-1}{\beta} \quad \text{or} \quad \beta \geq (1-\lambda)^{-1} \quad , \tag{10}$$

thus defining a critical value $\beta^c(\lambda) = (1-\lambda)^{-1}$. For $\beta \leq \beta^c(\lambda)$, the weight of compression is so strong that the solution degenerates to zero and no information is carried about X or Y. For $\beta \geq \beta^c(\lambda)$ the weight of information preservation is large enough, and the optimal solution for A is an eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$. The feasible regions for non degenerated solutions and the norm $||A||^2$ as a function of β and λ are depicted in Figure 2.

For some β values, several eigenvectors can satisfy the condition for non degenerated solutions of equation (10). Appendix C shows that the optimum is achieved by the eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ with the smallest eigenvalue. Note that this is also the eigenvector of $\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$ with the largest eigenvalue. We conclude that for scalar projections

$$A(\beta) = \begin{cases} \sqrt{\frac{\beta(1-\lambda)-1}{r\lambda}}v_1 & 0 < \lambda \leq \frac{\beta-1}{\beta} \\ 0 & \frac{\beta-1}{\beta} \leq \lambda \leq 1 \end{cases} \tag{11}$$

where v_1 is the eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ with the smallest eigenvalue.

4.2 The high-dimensional case

We now return to the proof of the general, high dimensional case, which follows the same lines as the scalar projection case. Setting the gradient in equation (7) to zero and reordering we obtain

$$\frac{\beta-1}{\beta}\left[(A\Sigma_{x|y}A^T+I_d)(A\Sigma_x A^T+I_d)^{-1}\right]A = A\left[\Sigma_{x|y}\Sigma_x^{-1}\right] \ . \tag{12}$$

Equation (12) shows that the multiplication of $\Sigma_{x|y}\Sigma_x^{-1}$ by A must reside in the span of the rows of A. This means that A should be spanned by up to n_t eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$. We can therefore represent the projection A as a mixture $A = WV$ where the rows of V are left normalized eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$ and W is a mixing matrix that weights these eigenvectors. The form of the mixing matrix W, that characterizes the norms of these eigenvectors, is described in the following lemma, which is proved in Appendix D.

Lemma 4.1 *The optimum of the cost function is obtained with a diagonal mixing matrix W of the form*

$$W = diag\left[\sqrt{\frac{\beta(1-\lambda_1)-1}{\lambda_1 r_1}};\ldots;\sqrt{\frac{\beta(1-\lambda_k)-1}{\lambda_k r_k}};0;\ldots;0\right] \tag{13}$$

where $\{\lambda_1,\ldots,\lambda_k\}$ are $k \le n_x$ eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ with critical β values ${\beta^c}_1,\ldots,{\beta^c}_k \le \beta$. $r_i \equiv \mathbf{v}_i^T\Sigma_x\mathbf{v}_i$ as in theorem 3.1.

The proof is presented in appendix D.

We have thus characterized the set of all minima of $\mathcal{L}$, and turn to identify which of them achieve the global minima.

Corollary 4.2
The global minimum of $\mathcal{L}$ is obtained with all λ_i that satisfy $\lambda_i < \frac{\beta-1}{\beta}$

The proof is presented in appendix D.

Taken together, these observations prove that for a given value of β, the optimal projection is obtained by taking all the eigenvectors whose eigenvalues λ_i satisfy $\beta \ge \frac{1}{1-\lambda_i}$, and setting their norm according to $A = WV$ with W determined as in Lemma 4.1. This completes the proof of Theorem 3.1.

5. The GIB Information Curve

The information bottleneck is targeted at characterizing the tradeoff between information preservation (accuracy of relevant predictions) and compression. Interestingly, much of the structure of the problem is reflected in the *information curve*, namely, the maximal value of relevant preserved information (accuracy), $I(T;Y)$, as function of the complexity of the representation of X, measured by $I(T;X)$. This curve is related to the rate-distortion function in lossy source coding, as well as to the achievability limit in source coding with side-information (Wyner, 1975, Cover and Thomas, 1991). It was shown to be concave under general conditions (Gilad-Bachrach et al., 2003), but its precise functional form depends on the joint distribution and can reveal properties of the hidden structure of the variables. Analytic forms for the information curve are known only for very special cases, such as Bernoulli variables and some intriguing self-similar distributions. The analytic characterization of the Gaussian IB problem allows us to obtain a closed form expression for the information curve in terms of the relevant eigenvalues.

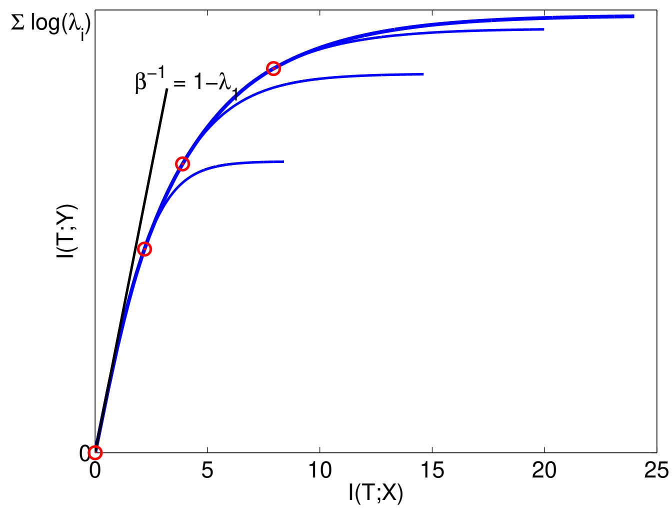

Figure 3: GIB information curve obtained with four eigenvalues $\lambda_i = 0.1,0.5,0.7,0.9$. The information at the critical points are designated by circles. For infinite β, curve is saturated at the log of the determinant $\sum \log \lambda_i$. For comparison, information curves calculated with smaller number of eigenvectors are also depicted (all curves calculated for $\beta < 1000$). The slope of the un-normalized curve at each point is the corresponding β^{-1}. The tangent at zero, with slope $\beta^{-1} = 1 - \lambda_1$, is super imposed on the information curve.

To this end, we substitute the optimal projection $A(\beta)$ into $I(T;X)$ and $I(T;Y)$ and rewrite them as a function of β

$$\begin{aligned} I_\beta(T;X) &= \frac{1}{2}\log\left(|A\Sigma_x A^T + I_d|\right) \qquad (14) \\ &= \frac{1}{2}\log\left(|(\beta(I-D)-I)D^{-1}|\right) \\ &= \frac{1}{2}\sum_{i=1}^{n(\beta)}\log\left((\beta-1)\frac{1-\lambda_i}{\lambda_i}\right) \\ I_\beta(T;Y) &= I(T;X) - \frac{1}{2}\sum_{i=1}^{n(\beta)}\log\beta(1-\lambda_i) \quad , \end{aligned}$$

where D is a diagonal matrix whose entries are the eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ as in appendix D, and $n(\beta)$ is the maximal index i such that $\beta \geq \frac{1}{1-\lambda_i}$. Isolating β as a function of $I_\beta(T;X)$

in the correct range of n_β and then $I_\beta(T;Y)$ as a function of $I_\beta(T;X)$ we have

$$I(T;Y) = I(T;X) - \frac{n_I}{2}\log\left(\prod_{i=1}^{n_I}(1-\lambda_i)^{\frac{1}{n_I}} + e^{\frac{2I(T;X)}{n_I}}\prod_{i=1}^{n_I}\lambda_i^{\frac{1}{n_I}}\right) \tag{15}$$

where the products are over the *first* $n_I = n_{\beta(I(T;X))}$ eigenvalues, since these obey the critical β condition, with $c_{n_I} \le I(T;X) \le c_{n_I+1}$ and $c_{n_I} = \sum_{i=1}^{n_I-1}\log\frac{\lambda_{n_I}}{\lambda_i}\frac{1-\lambda_i}{1-\lambda_{n_I}}$.

The GIB curve, illustrated in Figure 3, is continuous and smooth, but is built of several segments: as $I(T;X)$ increases additional eigenvectors are used in the projection. The derivative of the curve, which is equal to β^{-1}, can be easily shown to be continuous and decreasing, therefore the information curve is concave everywhere, in agreement with the general concavity of information curve in the discrete case (Wyner, 1975, Gilad-Bachrach et al., 2003). Unlike the discrete case where concavity proofs rely on the ability to use a large number of clusters, concavity is guaranteed here also for segments of the curve, where the number of eigenvectors are limited a-priori.

At each value of $I(T;X)$ the curve is bounded by a tangent with a slope $\beta^{-1}(I(T;X))$. Generally in IB, the data processing inequality yields an upper bound on the slope at the origin, $\beta^{-1}(0) < 1$, in GIB we obtain a tighter bound: $\beta^{-1}(0) < 1 - \lambda_1$. The asymptotic slope of the curve is always zero, as $\beta \to \infty$, reflecting the law of diminishing return: adding more bits to the description of X does not provide higher accuracy about T. This relation between the spectral properties of the covariance matrices raises interesting questions for special cases where the spectrum can be better characterized, such as random-walks and self-similar processes.

6. An iterative algorithm

The GIB solution is a set of scaled eigenvectors, and as such can be calculated using standard techniques. For example gradient ascent methods were suggested for learning CCA (Becker, 1996, Borga et al., 1997). An alternative approach is to use the general iterative algorithm for IB problems (Tishby et al., 1999). This algorithm that can be extended to continuous variables and representations, but its practical application for arbitrary distributions leads to a non-linear generalized eigenvalue problem whose general solution can be difficult. It is therefore interesting to explore the form that the iterative algorithm assumes once it is applied to Gaussian variables. Moreover, it may be possible to later extend this approach to more general parametric distributions, such as general exponential forms, for which linear eigenvector methods may no longer be adequate.

The general conditions for the IB stationary points were presented by Tishby et al. (1999) and can be written for a continuous variable x by the following self consistent equations for the unknown distributions $p(t|x)$, $p(y|t)$ and $p(t)$:

$$\begin{aligned} p(t) &= \int_X dx\, p(x)p(t|x) \\ p(y|t) &= \frac{1}{p(t)}\int_X dx\, p(x,y)p(t|x) \\ p(t|x) &= \frac{p(t)}{Z(\beta)}e^{-\beta D_{KL}[p(y|x)||p(y|t)]} \end{aligned} \tag{16}$$

where $Z(\beta)$ is a normalization factor (partition function) and is independent of x. It is important to realize that those conditions assume nothing about the representation variable T and should be satisfied by *any* fixed point of the IB Lagrangian. When X, Y and T have finite cardinality, those equations can be iterated directly in a Blahut-Arimoto like algorithm,

$$\begin{aligned} p(t_{k+1}|x) &= \frac{p(t_k)}{Z_{k+1}(x,\beta)} e^{-\beta D_{KL}[p(y|x)||p(y|t_k)]} \\ p(t_{k+1}) &= \int_X dx\, p(x) p(t_{k+1}|x) \\ p(y|t_{k+1}) &= \frac{1}{p(t_{k+1})} \int_X dx\, p(x,y) p(t_{k+1}|x)\ . \end{aligned} \tag{17}$$

where each iteration results in a distribution over the variables T_k, X and Y. The second and third equations calculate $p(t_{k+1})$ and $p(y|t_{k+1})$ using standard marginalization, and the Markov property $Y - X - T_k$. These iterations were shown to converge to the optimal T by Tishby et al. (1999).

For the general continuous T such an iterative algorithm is clearly not feasible. We show here, how the fact that we are confined to Gaussian distributions, can be used to turn those equations into an efficient parameter updating algorithm. We conjecture that algorithms for parameters optimizations can be defined also for parametric distribution other than Gaussians, such as other exponential distributions that can be efficiently represented with a small number of parameters.

In the case of Gaussian $p(x,y)$, when $p(t_k|x)$ is Gaussian for some k, so are $p(t_k)$, $p(y|t_k)$ and $p(t_{k+1}|x)$. In other words, the set of Gaussians $p(t|x)$ is invariant under the above iterations. To see why this is true, notice that $p(y|t_k)$ is Gaussian since T_k is jointly Gaussian with X. Also, $p(t_{k+1}|x)$ is Gaussian since $D_{KL}[p(y|x)||p(y|t_k)]$ between two Gaussians contains only second order moments in y and t and thus its exponential is Gaussian. This is in agreement with the general fact that the optima (which are fixed points of 17) are Gaussian (Globerson and Tishby, 2004). This invariance allows us to turn the IB algorithm that iterates over distributions, into an algorithm that iterates over the parameters of the distributions, being the relevant degrees of freedom in the problem.

Denote the variable T at time k by $T_k = A_k X + \xi_k$, where $\xi_k \sim \mathcal{N}(0, \Sigma_{\xi_k})$. The parameters A and Σ at time $k+1$ can be obtained by substituting T_k in the iterative IB equations. As shown in Appendix E, this yields the following update equations

$$\begin{aligned} \Sigma_{\xi_{k+1}} &= \left(\beta \Sigma_{t_k|y}^{-1} - (\beta-1)\Sigma_{t_k}^{-1}\right)^{-1} \\ A_{k+1} &= \beta \Sigma_{\xi_{k+1}} \Sigma_{t_k|y}^{-1} A_k \left(I - \Sigma_{y|x}\Sigma_x^{-1}\right) \end{aligned} \tag{18}$$

where $\Sigma_{t_k|y}, \Sigma_{t_k}$ are the covariance matrices calculated for the variable T_k.

This algorithm can be interpreted as repeated projection of A_k on the matrix $I - \Sigma_{y|x}\Sigma_x^{-1}$ (whose eigenvectors we seek) followed by scaling with $\beta \Sigma_{\xi_{k+1}} \Sigma_{t_k|y}^{-1}$. It thus has similar form to the power method for calculating the dominant eigenvectors of the matrix $\Sigma_{y|x}\Sigma_x^{-1}$ (Demmel, 1997, Golub and Loan, 1989). However, unlike the naive power method, where only the single dominant eigenvector is preserved, the GIB iterative algorithm maintains several

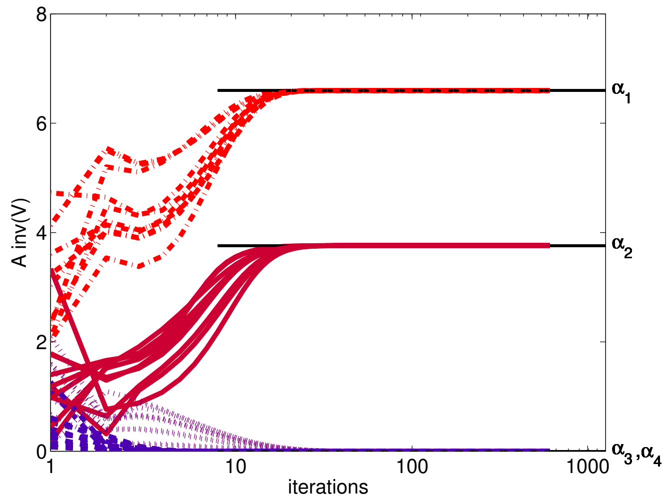

Figure 4: The norm of projection on the four eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$, as evolves along the operation of the iterative algorithm. Each line corresponds to the length of the projection of one row of A on the closest eigenvector. The projection on the other eigenvectors also vanishes (not shown). β was set to a value that leads to two non vanishing eigenvectors. The algorithm was repeated 10 times with different random initialization points, showing that it converges within 20 steps to the correct values α_i.

different eigenvectors, and their number is determined by the continuous parameter β and emerges from the iterations: All eigenvectors whose eigenvalues are smaller than the critical β vanish to zero, while the rest are properly scaled. This is similar to an extension of the naive power method known as *Orthogonal Iteration*, in which the projected vectors are renormalized to maintain several non vanishing vectors (Jennings and Stewart, 1975).

Figure 4 demonstrates the operation of the iterative algorithm for a four dimensional X and Y. The tradeoff parameter β was set to a value that leads to two vanishing eigenvectors. The norm of the other two eigenvectors converges to the correct values, which are given in Theorem 3.1.

The iterative algorithm can also be interpreted as a regression of X on T via Y. This can be seen by writing the update equation for A_{k+1} as

$$A_{k+1} = \Sigma_{\xi_{k+1}}\Sigma_{t_k|y}^{-1}\left(\Sigma_{yt_k}\Sigma_y^{-1}\right)\left(\Sigma_{yx}\Sigma_x^{-1}\right). \tag{19}$$

Since $\Sigma_{yx}\Sigma_x^{-1}$ describes the optimal linear regressor of X on Y, the operation of A_{k+1} on X can be described by the following diagram

$$X \xrightarrow{\Sigma_{yx}\Sigma_x^{-1}} \mu_{y|x} \xrightarrow{\Sigma_{yt_k}\Sigma_y^{-1}} \mu_{t_k|\mu_{y|x}} \xrightarrow{\Sigma_{\xi_{k+1}}\Sigma_{t_k|y}^{-1}} T_{k+1} \tag{20}$$

where the last step scales and normalizes T.

7. Relation To Other Works

7.1 Canonical correlation analysis and Imax

The GIB projection derived above uses weighted eigenvectors of the matrix $\Sigma_{x|y}\Sigma_x^{-1} = I - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$. Such eigenvectors are also used in *Canonical correlations Analysis* (CCA) (Hotelling, 1935, Thompson, 1984, Borga, 2001), a method of descriptive statistics that finds linear relations between two variables. Given two variables X, Y, CCA finds a set of basis vectors for each variable, such that the correlation coefficient between the projection of the variables on the basis vectors is maximized. In other words, it finds the bases in which the correlation matrix is diagonal and the correlations on the diagonal are maximized. The bases are the eigenvectors of the matrices $\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}$ and $\Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}$, and the square roots of their corresponding eigenvalues are the *canonical correlation coefficients*. CCA was also shown to be a special case of continuous Imax (Becker and Hinton, 1992, Becker, 1996), when the Imax networks are limited to linear projections.

Although GIB and CCA involve the spectral analysis of the same matrices, they have some inherent differences. First of all, GIB characterizes not only the eigenvectors but also their norm, in a way that that depends on the trade-off parameter β. Since CCA depends on the correlation coefficient between the compressed (projected) versions of X and Y, which is a *normalized* measure of correlation, it is invariant to a rescaling of the projection vectors. In contrast, for any value of β, GIB will choose one particular rescaling given by theorem 3.1.

While CCA is symmetric (in the sense that both X and Y are projected), IB is non symmetric and only the X variable is compressed. It is therefore interesting that both GIB and CCA use the same eigenvectors for the projection of X.

7.2 Multiterminal information theory

The Information Bottleneck formalism was recently shown (Gilad-Bachrach et al., 2003) to be closely related to the problem of source coding with side information (Wyner, 1975). In the latter, two *discrete* variables X, Y are encoded separately at rates R_x, R_y, and the aim is to use them to perfectly reconstruct Y. The bounds on the achievable rates in this case were found in (Wyner, 1975) and can be obtained from the IB information curve.

When considering continuous variables, lossless compression at finite rates is no longer possible. Thus, mutual information for continuous variables is no longer interpretable in terms of the actual number of encoding bits, but rather serves as an optimal measure of information between variables. The IB formalism, although coinciding with coding theorems in the discrete case, is more general in the sense that it reflects the tradeoff between compression and information preservation, and is not concerned with exact reconstruction.

Lossy reconstruction can be considered by introducing distortion measures as done for source coding of Gaussians with side information by Wyner (1978) and by Berger and Zamir (1999) (see also Pradhan, 1998), but these focus on the region of achievable rates under constrained distortion and are not relevant for the question of finding the representations which capture the information between the variables. Among these, the formalism closest to ours is that of Berger and Zamir (1999) where the distortion in reconstructing X is assumed to be small (high-resolution scenario). However, their results refer to encoding rates and as such go to infinity as the distortion goes to zero. They also analyze the problem for scalar Gaussian variables, but the one-dimensional setting does not reveal the interesting spectral properties and phase transitions which appear only in the multidimensional case discussed here.

7.3 Gaussian IB with side information

When handling real world data, the relevance variable Y often contains multiple structures that are correlated to X, although many of them are actually irrelevant. The information bottleneck with side information ($IBSI$) (Chechik and Tishby, 2002) alleviates this problem using side information in the form of an *irrelevance* variable Y^- about which information is removed. $IBSI$ thus aims to minimize

$$\mathcal{L} = I(X;T) - \beta\left(I(T;Y^+) - \gamma I(T;Y^-)\right) \tag{21}$$

This formulation can also be extended to the Gaussian case, in a manner similar to the original GIB functional. Looking at its derivative w.r.t. to the projection A yields

$$\begin{aligned}\frac{\delta\mathcal{L}}{\delta A} = (\quad & 1-\beta+\beta\gamma \quad)(A\Sigma_x A^T + I_d)^{-1}2A\Sigma_x \\ + \quad & \quad \beta \qquad\quad (A\Sigma_{x|y^+}A^T + I_d)^{-1}2A\Sigma_{x|y^+} \\ - \quad & \quad \beta\gamma \qquad\; (A\Sigma_{x|y^-}A^T + I_d)^{-1}2A\Sigma_{x|y^-} \ .\end{aligned}$$

While GIB relates to an eigenvalue problem of the form $\lambda A = A\Sigma_{x|y}\Sigma_x^{-1}$, GIB with side information ($GIBSI$) requires to solve of a matrix equation of the form $\lambda' A + \lambda^+ A\Sigma_{x|y^+}\Sigma_x^{-1} = \lambda^- A\Sigma_{x|y^-}\Sigma_x^{-1}$, which is similar in form to a generalized eigenvalue problem. However, unlike standard generalized eigenvalue problems, but as in the GIB case analyzed in this paper, the eigenvalues themselves depend on the projection A.

8. Practical implications

The GIB approach can be viewed as a method for finding the best linear projection of X, under a constraint on $I(T;X)$. Another straightforward way to limit the complexity of the projection is to specify its dimension in advance. Such an approach leaves open the question of the relative weighting of the resulting eigenvectors. This is the approach taken in classical CCA, where the number of eigenvectors is determined according to a statistical significance test, and their weights are then set to $\sqrt{1-\lambda_i}$. This expression is the correlation coefficient between the i^{th} CCA projections on X and Y, and reflects the amount of correlation captured by the i^{th} projection. The GIB weighting scheme is different, since it is derived to preserve maximum information under the compression constraint. To illustrate the difference, consider the case where $\beta = \frac{1}{1-\lambda_3}$, so that only two eigenvectors are used by GIB. The CCA scaling in this case is $\sqrt{1-\lambda_1}$, and $\sqrt{1-\lambda_2}$. The GIB weights are (up to a constant) $\alpha_1 = \sqrt{\frac{\lambda_3-\lambda_1}{\lambda_1 r_1}}, \alpha_2 = \sqrt{\frac{\lambda_3-\lambda_2}{\lambda_2 r_2}}$,, which emphasizes large gaps in the eigenspectrum, and can be very different from the CCA scaling.

This difference between CCA scaling and GIB scaling may have implications on two aspects of learning in practical applications. First, in applications involving compression of

Gaussian signals due to limitation on available band-width. This is the case in the growing field of sensor networks in which sensors are often very limited in their communication bandwidth due to energy constraints. In these networks, sensors communicate with other sensors and transmit information about their local measurements. For example, sensors can be used to monitor chemicals' concentrations, temperature or light conditions. Since only few bits can be transmitted, the information has to be compressed in a relevant way, and the relative scaling of the different eigenvectors becomes important (as in transform coding Goyal, 2001). As shown above, GIB describes the optimal transformation of the raw data into information conserving representation.

The second aspect where GIB becomes useful is in interpretation of data. Today, canonical correlation analysis is widely used for finding relations between multi-variate continuous variables, in particular in domains which are inherently high dimensional such as meteorology (von Storch and Zwiers, 1999) chemometrics (Antti et al., 2002) and functional MRI of brains (Friman et al., 2003). Since GIB weights the eigenvectors of the normalized cross correlation matrix in a different way than CCA, it may lead to very different interpretation of the relative importance of factors in these studies.

9. Discussion

We applied the information bottleneck method to continuous jointly Gaussian variables X and Y, with a continuous representation of the compressed variable T. We derived an analytic optimal solution as well as a new general algorithm for this problem (GIB) which is based solely on the spectral properties of the covariance matrices in the problem. The solutions for GIB are characterized in terms of the trade-off parameter β between compression and preserved relevant information, and consist of eigenvectors of the matrix $\Sigma_{x|y}\Sigma_x^{-1}$, continuously adding up vectors as more complex models are allowed. We provide an analytic characterization of the optimal tradeoff between the representation complexity and accuracy - the "information curve" - which relates the spectrum to relevant information in an intriguing manner. Besides its clean analytic structure, GIB offers a way for analyzing empirical multivariate data when only its correlation matrices can be estimated. In that case it extends and provides new information theoretic insight to the classical Canonical Correlation Analysis.

The most intriguing aspect of GIB is in the way the dimensionality of the representation changes with increasing complexity and accuracy, through the continuous value of the trade-off parameter β. While both mutual information values vary continuously on the smooth information curve, the dimensionality of the optimal projection T increases discontinuously through a cascade of structural (second order) phase transitions, and the optimal curve moves from one analytic segment to another. While this transition cascade is similar to the bifurcations observed in the application of IB to clustering through deterministic annealing, this is the first time such dimensional transitions are shown to exist in this context. The ability to deal with all possible dimensions in a single algorithm is a novel advantage of this approach compared to similar linear statistical techniques as CCA and other regression and association methods.

Interestingly, we show how the general IB algorithm which iterates over distributions, can be transformed to an algorithm that performs iterations over the distributions' *parame-*

ters. This algorithm, similar to multi-eigenvector power methods, converges to a solution in which the number of eigenvectors is determined by the parameter β, in a way that emerges from the iterations rather than defined a-priori.

For multinomial variables, the IB framework can be shown to be related in some limiting cases to maximum-likelihood estimation in a latent variable model (Slonim and Weiss, 2002). It would be interesting to see whether the GIB-CCA equivalence can be extended and give a more general understanding of the relation between IB and statistical latent variable models.

While the restriction to a Gaussian joint distribution deviates from the more general distribution independent approach of IB, it provides a precise example to the way representations with different dimensions can appear in the more general case. We believe that this type of dimensionality-transitions appears for more general distributions, as can be revealed in some cases by applying the Laplace method of integration (a Gaussian approximation) to the integrals in the general IB algorithm for continuous T.

The more general exponential forms, can be considered as a kernelized version of IB (see Mika et al., 2000) and appear in other minimum-information methods (such as SDR, Globerson and Tishby, 2003). these are of particular interest here, as they behave like Gaussian distributions in the joint kernel space. The Kernel Fisher-matrix in this case will take the role of the original cross covariance matrix of the variables in GIB.

Another interesting extension of our work is to networks of Gaussian processes. A general framework for that problem was developed in Friedman et al. (2001) and applied for discrete variables. In this framework the mutual information is replaced by multi-information, and the dependencies of the compressed and relevance variables is specified through two Graphical models. It is interesting to explore the effects of dimensionality changes in this more general framework, to study how they induce topological transitions in the related graphical models, as some edges of the graphs become important only beyond corresponding critical values of the tradeoff parameter β.

Acknowledgments

G.C. and A.G. were supported by the Israeli Ministry of Science, the Eshkol Foundation. This work was partially supported by a center of excellence grant of the Israeli Science Foundation (ISF).

Appendix A. Invariance to the noise covariance matrix

Lemma A.1 *For every pair (A, Σ_ξ) of a projection A and a full rank covariance matrix Σ_ξ, there exist a matrix $\tilde{A}$ such that $\mathcal{L}(\tilde{A}, I_d) = \mathcal{L}(A, \Sigma_\xi)$, where I_d is the $n_t \times n_t$ identity matrix.*

Proof: Denote by V the matrix which diagonalizes Σ_ξ, namely $\Sigma_\xi = VDV^T$, and by c the determinant $c \equiv |\sqrt{D^{-1}}V| = |\sqrt{D^{-1}}V^T|$. Setting $\tilde{A} \equiv \sqrt{D^{-1}}VA$ we have

$$\begin{aligned}
\mathcal{L}(\tilde{A}, I) &= (1-\beta)\log(|\tilde{A}\Sigma_x\tilde{A}^T + I_d|) - \log(|I_d|) + \beta\log(|\tilde{A}\Sigma_{x|y}\tilde{A}^T + I_d|) \\
&= (1-\beta)\log(c|A\Sigma_x A^T + \Sigma_\xi|c) - \log(c|\Sigma_\xi|c) + \beta\log(c|A\Sigma_{x|y}A^T + \Sigma_\xi|c) \\
&= (1-\beta)\log(|A\Sigma_x A^T + \Sigma_\xi|) - \log(|\Sigma_\xi|) + \beta\log(|A\Sigma_{x|y}A^T + \Sigma_\xi|) \\
&= \mathcal{L}(A, \Sigma_\xi)
\end{aligned} \tag{22}$$

where the first equality stems from the fact that the determinant of a matrix product is the product of the determinants. □

Appendix B. Properties of eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ and Σ_x

Lemma B.1 *Denote the set of left normalized eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$ by $\mathbf{v}_i$ ($||\mathbf{v}_i|| = 1$) and their corresponding eigenvalues by λ_i. Then*

1. *All the eigenvalues are real and satisfy $0 \le \lambda_i \le 1$*
2. *$\exists r_i > 0$ s.t. $\mathbf{v}_i^T \Sigma_x \mathbf{v}_j = \delta_{ij} r_i$.*
3. *$\mathbf{v}_i^T \Sigma_{x|y} \mathbf{v}_j = \delta_{ij}\lambda_i r_i$.*

The proof is standard (see e.g Golub and Loan, 1989) and is brought here for completeness.
Proof:

1. The matrices $\Sigma_{x|y}\Sigma_x^{-1}$ and $\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$ are positive semi definite (PSD), and their eigenvalues are therefore positive. Since $\Sigma_{x|y}\Sigma_x^{-1} = I - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$, the eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ are bounded between 0 and 1.

2. Denote by V the matrix whose rows are $\mathbf{v}_i^T$. The matrix $V\Sigma_x^{\frac{1}{2}}$ is the eigenvector matrix of $\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}}$ since $\left(V\Sigma_x^{\frac{1}{2}}\right)\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}} = V\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}} = \left(V\Sigma_{x|y}\Sigma_x^{-1}\right)\Sigma_x^{\frac{1}{2}} = DV\Sigma_x^{\frac{1}{2}}$. From the fact that $\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}}$ is symmetric, $V\Sigma_x^{\frac{1}{2}}$ is orthogonal, and thus $V\Sigma_x V^T$ is diagonal.

3. Follows from 2: $\mathbf{v}_i^T\Sigma_{x|y}\Sigma_x^{-1}\Sigma_x\mathbf{v}_j = \lambda_i\mathbf{v}_i^T\Sigma_x\mathbf{v}_j = \lambda_i\delta_{ij}r_i$.

□

Appendix C. Optimal eigenvector

For some β values, several eigenvectors can satisfy the conditions for non degenerated solutions (equation 10). To identify the optimal eigenvector, we substitute the value of $||A||^2$ from equation (9) $A\Sigma_{x|y}A^T = r\lambda||A||^2$ and $A\Sigma_x A^T = r||A||^2$ into the target function $\mathcal{L}$ of equation (6), and obtain

$$\mathcal{L} = (1-\beta)\log\left(\frac{(1-\lambda)(\beta-1)}{\lambda}\right) + \beta\log\left(\beta(1-\lambda)\right) \tag{23}$$

Since $\beta \geq 1$, this is monotonically increasing in λ and is minimized by the eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ with the smallest eigenvalue. Note that this is also the eigenvector of $\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$ with the largest eigenvalue.

Appendix D. Optimal mixing matrix

Lemma D.1 *The optimum of the cost function is obtained with a diagonal mixing matrix W of the form*

$$W = diag\left[\sqrt{\frac{\beta(1-\lambda_1)-1}{\lambda_1 r_1}};\ldots;\sqrt{\frac{\beta(1-\lambda_k)-1}{\lambda_k r_k}};0;\ldots;0\right] \tag{24}$$

where $\{\lambda_1,\ldots,\lambda_k\}$ are $k \leq n_x$ eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ with critical β values ${\beta^c}_1,\ldots,{\beta^c}_k \leq \beta$. $r_i \equiv \mathbf{v}_i^T\Sigma_x\mathbf{v}_i$ as in theorem 3.1.

Proof: We write $V\Sigma_{x|y}\Sigma_x^{-1} = DV$ where D is a diagonal matrix whose elements are the corresponding eigenvalues, and denote by R the diagonal matrix whose i^{th} element is r_i. When $k = n_x$, we substitute $A = WV$ into equation (12), and eliminate V from both sides to obtain

$$\frac{\beta-1}{\beta}\left[(WDRW^T + I_d)(WRW^T + I_d)^{-1}\right]W = WD \tag{25}$$

Use the fact that W is full rank to multiply by W^{-1} from the left and by $W^{-1}(WRW^T + I_d)W$ from the right

$$\frac{\beta-1}{\beta}(DRW^TW + I_d) = D(RW^TW + I_d) \tag{26}$$

Rearranging, we have,

$$W^TW = [\beta(I-D) - I](DR)^{-1} \tag{27}$$

which is a diagonal matrix.

While this does not uniquely characterize W, we note that using properties of the eigenvalues from lemma B.1, we obtain

$$|A\Sigma_x A^T + I_d| \quad = \quad |WV\Sigma_x V^T W^T + I_d| = |WRW^T + I_d| \ .$$

Note that WRW^T has left eigenvectors W^T with corresponding eigenvalues obtained from the diagonal matrix W^TWR. Thus if we substitute A into the target function in equation (6), a similar calculation yields

$$\mathcal{L} = (1-\beta)\sum_{i=1}^{n}\log\left(||\mathbf{w}_i^T||^2 r_i + 1\right) + \beta\sum_{i=1}^{n}\log\left(||\mathbf{w}_i^T||^2 r_i\lambda_i + 1\right) \tag{28}$$

where $||\mathbf{w}_i^T||^2$ is the i^{th} element of the diagonal of W^TW. This shows that $\mathcal{L}$ depends only on the norm of the columns of W, and all matrices W that satisfy (27) yield the same target function. We can therefore choose to take W to be the diagonal matrix which is the (matrix) square root of (27)

$$W = \sqrt{[\beta(I-D)-I](DR)^{-1}} \tag{29}$$

which completes the proof of the full rank ($k = n_x$) case.

In the low rank ($k < n_x$) case W does not mix all the eigenvectors, but only k of them. To prove the lemma for this case, we first show that any such low rank matrix is equivalent (in terms of the target function value) to a low rank matrix that has only k non zero rows. We then conclude that the non zero rows should follow the form described in the above lemma.

Consider a $n_x \times n_x$ matrix W of rank $k < n_x$, but without any zero rows. Let U be the set of left eigenvectors of WW^T (that is, $WW^T = U\Lambda U^T$). Then, since WW^T is Hermitian, its eigenvectors are orthonormal, thus $(UW)(WU)^T = \Lambda$ and $W' = UW$ is a matrix with k non zero rows and $n_x - k$ zero lines. Furthermore, W' obtains the same value of the target function, since

$$\begin{aligned}\mathcal{L} &= (1-\beta)\log(|W'RW'^T + \Sigma_\xi^2|) + \beta\log(|W'DRW'^T + \Sigma_\xi^2|) \\ &= (1-\beta)\log(|UWRW^TU^T + UU^T\Sigma_\xi^2|) + \beta\log(|UWDRW^TU^T + UU^T\Sigma_\xi^2|) \\ &= (1-\beta)\log(|U||WRW^T + \Sigma_\xi^2||U^T|) + \beta\log(|U||UWDRW^TU^T + \Sigma_\xi^2||U^T|) \\ &= (1-\beta)\log(|WRW^T + \Sigma_\xi^2|) + \beta\log(|WDRW^TT + \Sigma_\xi^2|)\end{aligned} \tag{30}$$

where we have used the fact that U is orthonormal and hence $|U| = 1$. To complete the proof note that the non zero rows of W' also have $n_x - k$ zero columns and thus define a square matrix of rank k, for which the proof of the full rank case apply, but this time by projecting to a dimension k instead of n_x. $\square$

This provides a characterization of all local minima. To find which is the global minimum, we prove the following corollary.

Corollary D.2
The global minimum of $\mathcal{L}$ is obtained with all λ_i that satisfy $\lambda_i < \frac{\beta-1}{\beta}$

Proof: Substituting the optimal W of equation (29) into equation (28) yields

$$\mathcal{L} = \sum_{i=1}^{k}(\beta-1)\log\lambda_i + \log(1-\lambda_i) + f(\beta). \tag{31}$$

Since $0 \leq \lambda \leq 1$ and $\beta \geq \frac{1}{1-\lambda}$, $\mathcal{L}$ is minimized by taking all the eigenvalues that satisfy $\beta > \frac{1}{(1-\lambda_i)}$. $\square$

Appendix E. Deriving the iterative algorithm

To derive the iterative algorithm in section 6, we assume that the distribution $p(t_k|x)$ corresponds to the Gaussian variable $T_k = A_k X + \xi_k$. We show below that $p(t_{k+1}|x)$ corresponds to $T_{k+1} = A_{k+1}X + \xi_{k+1}$ with $\xi_{k+1} \sim N(0, \Sigma_{\xi_{k+1}})$ and

$$\begin{aligned} \Sigma_{\xi_{k+1}} &= \left(\beta\Sigma^{-1}_{t_k|y} - (\beta-1)\Sigma^{-1}_{t_k}\right)^{-1} \\ A_{k+1} &= \beta\Sigma_{\xi_{k+1}}\Sigma^{-1}_{t_k|y}A_k\left(I - \Sigma_{y|x}\Sigma^{-1}_x\right) \end{aligned} \tag{32}$$

We first substitute the Gaussian $p(t_k|x) \sim N(A_k x, \Sigma_{\xi_k})$ into the equations of (17), and treat the second and third equations. The second equation $p(t_k) = \int_x p(x)p(t_k|x)dx$, is a marginal of the Gaussian $T_k = A_k X + \xi_k$, and yields a Gaussian $p(t_k)$ with zero mean and covariance

$$\Sigma_{t_k} = A_k\Sigma_x A_k^T + \Sigma_{\xi_k} \tag{33}$$

The third equation, $p(y|t_k) = \frac{1}{p(t_k)}\int_x p(x,y)p(t_k|x)dx$ defines a Gaussian with mean and covariance matrix given by:

$$\begin{aligned} \mu_{y|t_k} &= \mu_y + \Sigma_{yt_k}\Sigma^{-1}_{t_k}(t_k - \mu_{t_k}) = \Sigma_{yt_k}\Sigma^{-1}_{t_k}t_k \equiv B_k t_k \\ \Sigma_{y|t_k} &= \Sigma_y - \Sigma_{yt_k}\Sigma^{-1}_{t_k}\Sigma_{t_k y} = \Sigma_y - A_k\Sigma_{xy}\Sigma^{-1}_{t_k}\Sigma_{yx}A_k^T \end{aligned} \tag{34}$$

where we have used the fact that $\mu_y = \mu_{t_k} = 0$, and define the matrix $B_k \equiv \Sigma_{yt_k}\Sigma^{-1}_{t_k}$ as the regressor of t_k on y. Finally, we return to the first equation of (17), that defines $p(t_{k+1}|x)$ as

$$p(t_{k+1}|x) = \frac{p(t_k)}{Z(x,\beta)}e^{-\beta D_{KL}[p(y|x)||p(y|t_k)]} \tag{35}$$

We now show that $p(t_{k+1}|x)$ is Gaussian and compute its mean and covariance matrix.

The KL divergence between the two Gaussian distributions, in the exponent of equation (35) is known to be

$$\begin{aligned} 2D_{KL}[p(y|x)||p(y|t_k)] &= \log\frac{|\Sigma_{y|t_k}|}{|\Sigma_{y|x}|} + Tr(\Sigma^{-1}_{y|t_k}\Sigma_{y|x}) \\ &+ (\mu_{y|x} - \mu_{y|t_k})^T\Sigma^{-1}_{y|t_k}(\mu_{y|x} - \mu_{y|t_k}) \end{aligned} \tag{36}$$

The only factor which explicitly depends on the value of t in the above expression is $\mu_{y|t_k}$ derived in equation (34), is linear in t. The KL divergence can thus be rewritten as

$$D_{KL}[p(y|x)||p(y|t_k)] = c(x) + \frac{1}{2}(\mu_{y|x} - B_k t_k)^T\Sigma^{-1}_{y|t_k}(\mu_{y|x} - B_k t_k)$$

Adding the fact that $p(t_k)$ is Gaussian we can write the log of equation (35) as a quadratic form in t

$$\log p(t_{k+1}|x) = Z(x) + (t_{k+1} - \mu_{t_{k+1}|x})^T\Sigma_{\xi_{k+1}}(t_{k+1} - \mu_{t_{k+1}|x})$$

where

$$\begin{aligned} \Sigma_{\xi_{k+1}} &= \left(\beta B_k^T \Sigma_{y|t_k}^{-1} B_k + \Sigma_{t_k}^{-1}\right)^{-1} \\ \mu_{t_{k+1}|x} &= A_{k+1} x \\ A_{k+1} &= \beta \Sigma_{\xi_{k+1}} B_k^T \Sigma_{y|t_k}^{-1} \Sigma_{yx} \Sigma_x^{-1} x \end{aligned} \tag{37}$$

This shows that $p(t_{k+1}|x)$ is a Gaussian $T_{k+1} = A_{k+1}x + \xi_{k+1}$, with $\xi \sim N(0, \Sigma_{\xi_{k+1}})$.

To simplify the form of $A_{k+1}, \Sigma_{\xi_{k+1}}$, we use the two following matrix inversion lemmas [3], which hold for any matrices E, F, G, H of appropriate sizes when E, H are invertible.

$$\begin{aligned} (E - FH^{-1}G)^{-1} &= E^{-1} + E^{-1}F(H - GE^{-1}F)^{-1}GE^{-1} \\ (E - FH^{-1}G)^{-1}FH^{-1} &= E^{-1}F(H - GE^{-1}F)^{-1} \end{aligned} \tag{38}$$

Using $E \equiv \Sigma_{t_k}$, $F \equiv \Sigma_{yt_k}$, $H \equiv \Sigma_y$, $G \equiv \Sigma_{yt_k}$, $B_k = \Sigma_{yt_k}\Sigma_{t_k}^{-1}$ in the first lemma we obtain

$$\Sigma_{t_k|y}^{-1} = \Sigma_{t_k}^{-1} + B_k^T \Sigma_{y|t_k}^{-1} B_k \quad .$$

Replacing this into the expression for $\Sigma_{\xi_{k+1}}$ in equation (37) we obtain

$$\Sigma_{\xi_{k+1}} = \left(\beta \Sigma_{t_k|y}^{-1} - (\beta - 1)\Sigma_{t_k}^{-1}\right)^{-1} \tag{39}$$

Finally, using again $E \equiv \Sigma_{t_k}$, $F \equiv \Sigma_{t_k y}$, $H \equiv \Sigma_y$, $G \equiv \Sigma_{yt_k}$ in the second matrix lemma, we have $\Sigma_{t_k|y}^{-1}\Sigma_{t_k y}\Sigma_y^{-1} = \Sigma_{t_k}^{-1}\Sigma_{t_k y}\Sigma_{y|t_k}^{-1}$, which turns the expression for A_{k+1} in equation (37) into

$$\begin{aligned} A_{k+1} &= \beta \Sigma_{\xi_{k+1}} \Sigma_{t_k|y}^{-1} \Sigma_{t_k y} \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \\ &= \beta \Sigma_{\xi_{k+1}} \Sigma_{t_k|y}^{-1} A_k \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \\ &= \beta \Sigma_{\xi_{k+1}} \Sigma_{t_k|y}^{-1} A_k (I - \Sigma_{x|y} \Sigma_x^{-1}) \ , \end{aligned} \tag{40}$$

which completes the derivation of the algorithm as described in (17).

References

H. Antti, E. Holmes, and J. Nicholson. Multivariate solutions to metabonomic profiling and functional genomics. part 2 - chemometric analysis, 2002. http://www.acc.umu.se/ tnkjtg/chemometrics/editorial/oct2002.

S. Becker. Mutual information maximization: Models of cortical self-organization. *Network: Computation in Neural Systems*, pages 7,31, 1996.

S. Becker and G.E. Hinton. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

3. The first equation is the standard inversion lemma (see e.g., Kay, 1993, page 571). The second can be easily verified from the first.

T. Berger and R. Zamir. A semi-continuous version of the berger-yeung problem. *IEEE Transactions on Information Theory*, pages 1520–1526, 1999.

M. Borga. Canonical correlation: a tutorial. http://people.imt.liu.se/magnus/cca, January 2001.

M. Borga, H. Knutsson, and T. Landelius. Learning canonical correlations. In *Proceedings of the 10th Scandinavian Conference on Image Analysis*, Lappeenranta, Finland, June 1997. SCIA.

G. Chechik and N. Tishby. Extracting relevant structures with side information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, Vancouver, Canada, 2002.

T.M. Cover and J.A. Thomas. *The elements of information theory*. Plenum Press, New York, 1991.

J.W. Demmel. *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics, Philadelphia, 1997. ISBN 0-89871-389-7.

Alexander G. Dimitrov and John P. Miller. Neural coding and decoding: Communication channels and quantization. *Network: Computation in Neural Systems*, 12(4):441–472, 2001.

N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In J.S. Breese and D. Koller, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference (UAI-2001)*, pages 152–161, San Francisco, CA, 2001. Morgan Kaufmann.

O. Friman, M. Borga, P. Lundberg, and H. Knutsson. Adaptive analysis of fMRI data. *NeuroImage*, 19(3):837–845, 2003.

Rimoldi Gastpar and Vetterli. To code, or not to code: lossy source-channel communication revisited. *Information Theory*, 49(5):1147–1158, 2003.

R. Gilad-Bachrach, A. Navot, and N. Tishby. An information theoretic tradeoff between complexity and accuracy. In *Proceedings of the COLT*, Washington, 2003.

A. Globerson and N. Tishby. Sufficient dimensionality reduction. *Journal of Macine Learning Research*, 3:1307–1331, 2003. ISSN 1533-7928.

A. Globerson and N. Tishby. Tbd. Technical report, Hebrew University, January 2004.

G.H. Golub and C.F.V. Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 1989.

V.K. Goyal. Theoretical foundations of transform coding. *Signal Processing Magazine, IEEE*, 18(5):9–21, 2001.

H. Hotelling. The most predictable criterion. *Journal of Educational Psychology,*, 26:139–142, 1935.

A. Jennings and G.W. Stewart. Simultaneous iteration for partial eigensolution of real matrices. *J. Inst. Math Appl*, 15:351–361, 1975.

S.M. Kay. *Fundamentals of Statistical Signal Processing. Volume I, Estimation Theory.* Prentice-Hall, 1993.

R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

R. Linsker. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation*, 4:691–702, 1992.

J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley and Sons, 1988.

S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. Smola, and K. Muller. Invariant feature extraction and classification in kernel spaces. In S.A. Solla, T.K. Leen, and K.R. Muller, editors, *Advances in Neural Information Processing Systems 12*, pages 526–532, Vancouver, Canada, 2000.

S.S. Pradhan. On rate-distortion function of gaussian sources with memory with side information at the decoder. Technical report, Berkeley, 1998.

S.S. Pradhan, J. Chou, and K. Ramchandran. Duality between source coding and channel coding and its extension to the side information case. *Information Theory*, 49(5):1181–1203, 2003.

Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *Institute for Radio Engineers, International Convention Record*, volume 7, part 4, pages 142–163, New York, NY, USA, March 1959.

J. Sinkkonen and S. Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2001.

N. Slonim. *Information Bottlneck theory and applications.* PhD thesis, Hebrew University of Jerusalem, 2003.

N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In P. Ingwersen N.J. Belkin and M-K. Leong, editors, *Research and Development in Information Retrieval (SIGIR-00)*, pages 208–215. ACM press, new york, 2000.

N. Slonim and Y. Weiss. Maximum likelihood and the information bottleneck. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, Vancouver, Canada, 2002.

B. Thompson. *Canonical correlation analysis: Uses and interpretation.*, volume 47. Thousands Oak, CA Sage publications, 1984.

N. Tishby, F.C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of 37th Allerton Conference on communication and computation*, 1999.

H. von Storch and F.W. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, 1999.

A.D. Wyner. On source coding with side information at the decoder. *IEEE Trans. on Info Theory*, IT-21:294–300, 1975.

A.D. Wyner. The rate distortion function for source coding with side information at the decoder ii: General sources. *IEEE Trans. on Info Theory*, IT-38:60–80, 1978.