Auxiliary Variational Information Maximization for Dimensionality Reduction

Felix Agakov[1] and David Barber[2]

[1] University of Edinburgh, 5 Forrest Hill, EH1 2QL Edinburgh, UK
felixa@inf.ed.ac.uk, www.anc.ed.ac.uk
[2] IDIAP, Rue du Simplon 4, CH-1920 Martigny Switzerland,
www.idiap.ch

Abstract. Mutual Information (MI) is a long studied measure of information content, and many attempts to apply it to feature extraction and stochastic coding have been made. However, in general MI is computationally intractable to evaluate, and most previous studies redefine the criterion in forms of approximations. Recently we described properties of a simple lower bound on MI, and discussed its links to some of the popular dimensionality reduction techniques. Here we introduce a richer family of *auxiliary variational* bounds on MI, which generalizes our previous approximations. Our specific focus then is on applying the bound to extracting informative lower-dimensional projections in the presence of irreducible Gaussian noise. We show that our method produces significantly tighter bounds than the well-known *as-if Gaussian* approximations of MI. We also show that the auxiliary variable method may help to significantly improve on reconstructions from noisy lower-dimensional projections. Interestingly, it may be shown that our information-theoretic approach to stochastic dimensionality reduction generalizes self-supervised training in stochastic autoencoders.

1 Introduction

One of the principal goals of dimensionality reduction is to produce a lower-dimensional representation y of a high-dimensional source vector x, so that the useful information contained in the source data is not lost. If it is not known a priori which coordinates of x may be relevant for a specific task, it is sensible to maximize the amount of information which y contains about all the coordinates, for all possible x's. The fundamental measure in this context is the mutual information

$$I(\mathsf{x},\mathsf{y}) \equiv H(\mathsf{x}) - H(\mathsf{x}|\mathsf{y}), \tag{1}$$

which indicates the decrease of uncertainty in x due to the knowledge of y. Here $H(\mathsf{x}) \equiv -\langle \log p(\mathsf{x}) \rangle_{p(\mathsf{x})}$ and $H(\mathsf{x}|\mathsf{y}) \equiv -\langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x},\mathsf{y})}$ are marginal and conditional entropies respectively, and the angled brackets represent averages over all variables contained within the brackets.

The principled information theoretic approach to dimensionality reduction maximizes (1) with respect to parameters of the *encoder* $p(\mathsf{y}|\mathsf{x})$. However, it is

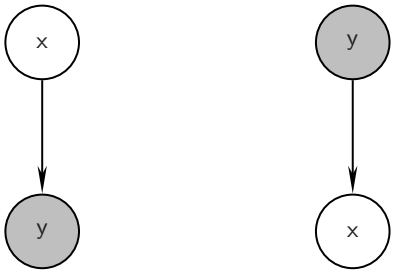


Fig. 1. Generative and encoder models. (a): An encoder model $\mathcal{M}_I \stackrel{\text{def}}{=} \tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})$ trained by maximizing the mutual information $I(\mathsf{x},\mathsf{y})$ (b): a generative model $\mathcal{M}_L \stackrel{\text{def}}{=} p(\mathsf{y})p(\mathsf{x}|\mathsf{y})$ trained by maximizing the log-likelihood $\langle \log p(\mathsf{x}) \rangle_{\tilde{p}(\mathsf{x})}$. Here $\tilde{p}(\mathsf{x})$ is the empirical distribution, the shaded nodes indicate the hidden representations y, and we have assumed that the data patterns x are i.i.d.

easy to see that if the reduced dimension $|\mathsf{y}|$ ($|\mathsf{y}| < |\mathsf{x}|$) is still large, the exact evaluation of $I(\mathsf{x},\mathsf{y})$ is in general computationally intractable. The key difficulty lies in the computation of the conditional entropy $H(\mathsf{x}|\mathsf{y})$, which is tractable only in a few special cases. Typically, the standard techniques assume that $p(\mathsf{x},\mathsf{y})$ is jointly Gaussian (so that $I(\mathsf{x},\mathsf{y})$ has a closed analytical form), or the channels are deterministic and invertible [12], [5] (which may be related to the noiseless square ICA case). Alternatively, it is sometimes assumed that the output spaces are very low-dimensional, so that integration over y in the computation of $I(\mathsf{x},\mathsf{y})$ may be performed numerically). Unfortunately, these assumptions may be too restrictive for many practical applications of subspace selection. Other existing methods suggest to optimize alternative objective functions (e.g. approximations of $I(\mathsf{x},\mathsf{y})$ based on the *Fisher Information* criterion [8]), which, however, do not retain proper bounds on $I(\mathsf{x},\mathsf{y})$ and may often lead to numerical instabilities when applied to learning undercomplete representations [1].

1.1 Encoder *vs* Generative Models

A principal motivation for applying information theoretic techniques for stochastic subspace selection and dimensionality reduction is the general intuition that the unknown compressed representations should be predictive about the higher-dimensional data. Additionally, we note that the information-maximizing framework of encoder models is particularly convenient for addressing problems of *constrained* dimensionality reduction, as by parameterizing the model $p(\mathsf{y}|\mathsf{x})$ we may easily impose specific parametric constraints on the possibly noisy projection to a lower-dimensional space (see Fig. 1 (a)). This is in contrast with

generative latent variable models (Fig. 1 (b)) commonly used for probabilistic dimensionality reduction (e.g. [7], [14], [16]), where the probabilistic projection to the lower-dimensional space $p(\mathsf{y}|\mathsf{x}) \propto p(\mathsf{y})p(\mathsf{x}|\mathsf{y})$ is a functional of the explicitly parameterized prior $p(\mathsf{y})$ and the generating conditional $p(\mathsf{x}|\mathsf{y})$. Effectively, parameterizing an encoder model is analogous to specifying a conditionally trained discriminative regressor; however, in contrast to discriminative models, the lower-dimensional vectors y will in our case be hidden. Finally, we note that training encoder models by optimizing the likelihood is effectively meaningless, as the unknown representations y would marginalize out. On the other hand, training such models by maximizing the mutual information (1) will generally require approximations.

1.2 Linsker's *as-if Gaussian* approximation

A popular class of methods suggests to approximate $I(\mathsf{x}, \mathsf{y})$ by assuming that the joint distribution $p(\mathsf{x})p(\mathsf{y}|\mathsf{x}) \approx p_G(\mathsf{x}, \mathsf{y}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a Gaussian, independently of the exact form of $p(\mathsf{x}, \mathsf{y})$ [13]. Note that the conditional entropy $H(\mathsf{x}|\mathsf{y})$ may in this case be approximated by

$$H_G(\mathsf{x}|\mathsf{y}) \stackrel{\text{def}}{=} -\langle \log p_G(\mathsf{x}|\mathsf{y}) \rangle_{p_G(\mathsf{x},\mathsf{y})} = (1/2) \log(2\pi e)^{|\mathsf{x}|} |\boldsymbol{\Sigma}_{\mathsf{x}|\mathsf{y}}|, \tag{2}$$

where $\boldsymbol{\Sigma}_{\mathsf{x}|\mathsf{y}}$ is the covariance of the decoder $p_G(\mathsf{x}|\mathsf{y})$ expressed from $p_G(\mathsf{x}, \mathsf{y}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If the joint covariance is expressed as $\boldsymbol{\Sigma} \stackrel{\text{def}}{=} \langle [\mathsf{x}\ \mathsf{y}][\mathsf{x}\ \mathsf{y}]^T \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} - \langle [\mathsf{x}\ \mathsf{y}] \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} \langle [\mathsf{x}\ \mathsf{y}]^T \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}$, it is easy to obtain the *as-if Gaussian* approximation of $I(\mathsf{x}, \mathsf{y})$:

$$I_G(\mathsf{x}, \mathsf{y}) \propto \log |\boldsymbol{\Sigma}_{xx}| - \log |\boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{xy}^T|. \tag{3}$$

Here $\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{xy}$, and $\boldsymbol{\Sigma}_{yy}$ are the partitions of $\boldsymbol{\Sigma}$, and $\tilde{p}(\mathsf{x})$ is the empirical distribution. Objective (3) is to be maximized with respect to parameters of the encoder distribution $p(\mathsf{y}|\mathsf{x})$. After training, the encoder may be used for generating lower-dimensional representations y for a given source x. Inference is simpler than that in generative models and does not require additional evaluations.

2 A Simple Variational Lower Bound on $I(\mathsf{x}, \mathsf{y})$

In [4] we discussed properties of a simple variational lower bound on the mutual information $I(\mathsf{x}, \mathsf{y})$. The bound follows from non-negativity of the Kullback-Leibler divergence $KL(p(\mathsf{x}|\mathsf{y})||q(\mathsf{x}|\mathsf{y}))$ between the exact posterior $p(\mathsf{x}|\mathsf{y})$ and its variational approximation $q(\mathsf{x}|\mathsf{y})$, leading to

$$I(\mathsf{x}, \mathsf{y}) \geq \tilde{I}(\mathsf{x}, \mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x},\mathsf{y})}, \tag{4}$$

where $q(\mathsf{x}|\mathsf{y})$ is an arbitrary distribution. Clearly, the bound is saturated for $q(\mathsf{x}|\mathsf{y}) \equiv p(\mathsf{x}|\mathsf{y})$; however, in general this choice would lead to intractability of learning the optimal encoder $p(\mathsf{y}|\mathsf{x})$.

Objective (4) explicitly includes both the encoder $p(\mathsf{y}|\mathsf{x})$ (distribution of the lower-dimensional representations for a given source) and decoder $q(\mathsf{x}|\mathsf{y})$ (reconstruction of the source from a given compressed representation). It is iteratively optimized for parameters of both distributions (the *IM* algorithm [4]), which is qualitatively similar to the variational expectation-maximizing algorithm for intractable generative models. (Note, however, that optimization surfaces defined by the objectives of the IM and the variational EM are quite different). The flexibility in the choice of the decoder $q(\mathsf{x}|\mathsf{y})$ makes (4) particularly computationally convenient. Specifically, we may avoid most of the computational difficulties of optimizing $I(\mathsf{x},\mathsf{y})$ by constraining $q(\mathsf{x}|\mathsf{y})$ to lie in a tractable family. Note that the fundamental difference of optimizing the variational bound (4) from the well-known family of Blahut-Arimoto algorithms for channel capacity (e.g. [10]) is exactly the fact that the variational decoder distribution $q(\mathsf{x}|\mathsf{y})$ is constrained to be tractable. For example, $q(\mathsf{x}|\mathsf{y})$ may be chosen to have a simple parametric form or a sparse structure. Such constraints significantly facilitate optimization of channel capacity for non-trivial stochastic projections[3].

It is easy to show that by constraining the decoder as $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}(\mathsf{U}\mathsf{y}, \boldsymbol{\Sigma})$, optimization of the bound (4) reduces to maximization of Linsker's *as-if Gaussian* criterion (3). Therefore, maximization of I_G may be seen as a special case of the variational information-maximization approach for the case when the decoder $q(\mathsf{x}|\mathsf{y})$ is a linear Gaussian. Moreover, if for this case $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}\left(\mathsf{W}(\mathsf{x} - \langle\mathsf{x}\rangle_{\tilde{p}(\mathsf{x})}), s^2\mathsf{I}\right)$, it is easy to show that the left singular vectors of the optimal projection weights W^T correspond to the $|\mathsf{y}|$-PCA solution on the sample covariance $\langle\mathsf{x}\mathsf{x}^T\rangle_{\tilde{p}(\mathsf{x})} - \langle\mathsf{x}\rangle_{\tilde{p}(\mathsf{x})}\langle\mathsf{x}\rangle^T_{\tilde{p}(\mathsf{x})}$.

3 An Auxiliary Variational Bound

A principal conceptual difficulty of applying the bound (4) is in specifying a powerful yet tractable variational decoder $q(\mathsf{x}|\mathsf{y})$. Specifically, for isotropic Gaussian channels, the linear Gaussian decoders mentioned above are fundamentally limited to producing PCA projections. Here we describe a richer family of bounds on $I(\mathsf{x},\mathsf{y})$ which helps to overcome this limitation.

From (4) it is intuitive that we may obtain tighter bounds on $I(\mathsf{x},\mathsf{y})$ by increasing representational power of the variational distributions $q(\mathsf{x}|\mathsf{y})$. One way to achieve this is to consider multi-modal decoders $q(\mathsf{x}|\mathsf{y}) = \langle q(\mathsf{x}|\mathsf{y},\mathsf{z})\rangle_{q(\mathsf{z}|\mathsf{y})}$, where the introduced *auxiliary variables* z are effectively the unknown mixture states. Effectively, this choice of the variational decoder has a structure of a constrained multi-dimensional mixture-of-experts [11] model of a conditional distribution. Clearly, the fully-coupled structure of the resulting variational distribution $q(\mathsf{x}|\mathsf{y})$ qualitatively agrees with the structure of the exact posterior, as

[3] Standard iterative approaches to maximizing $I(\mathsf{x},\mathsf{y})$ in encoder models require optimization of the cross-entropy $\langle\log p^{(old)}(\mathsf{y})\rangle_{p(\mathsf{y})}$ between two fully-coupled distributions $p(\mathsf{y})$ and $p^{(old)}(\mathsf{y})$ for $p(\mathsf{y}|\mathsf{x})$ (see [3], [6], [10]), which is rarely tractable in practice.

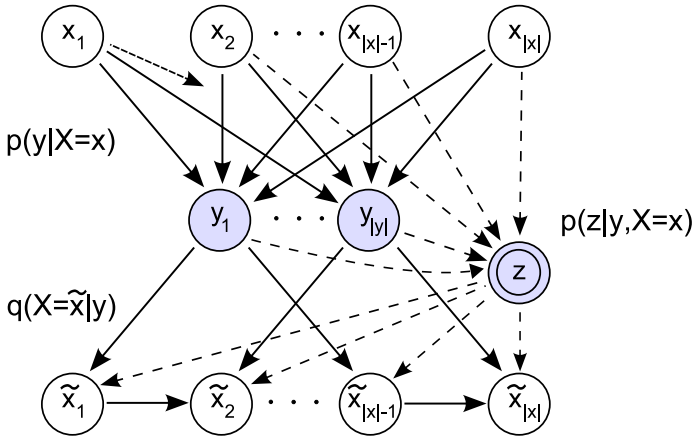


Fig. 2. A stochastic channel $p(\mathsf{y}|\mathsf{x})$ with a structured mixture-type decoder $q(\mathsf{x}|\mathsf{y})$. (The states of the reconstructed variables are denoted by $\tilde{\mathsf{x}}$). The auxiliary variables z are *not* transmitted across the channel $p(\mathsf{y}|\mathsf{x})$ and do not explicitly constrain $p(\mathsf{x},\mathsf{y})$. The dashed lines show the mappings to and from the auxiliary space. The auxiliary nodes are shown by the double circle.

different dimensions of the reconstructed vectors x are coupled through the auxiliary variables z. Moreover, for any interesting choice of the auxiliary space $\{\mathsf{z}\}$, the decoder $q(\mathsf{x}|\mathsf{y})$ will typically be multi-modal, which agrees with the generally multi-modal form of Bayesian decoders $p(\mathsf{x}|\mathsf{y})$. We may therefore intuitively hope that this choice of the variational posterior will generally result in tighter bounds on $I(\mathsf{x},\mathsf{y})$.

A possible disadvantage of mixture decoders $q(\mathsf{x}|\mathsf{y})$ relates to the fact that specifying the conditional mixing coefficients $q(\mathsf{z}|\mathsf{y})$ in a principled manner may be rather difficult. Moreover, if the auxiliary variables z are independent from the *original* source patterns x given the lower-dimensional encodings y, any noise in y will affect determining of the mixing states. Intuitively, this may have an overwhelming negative effect on decoding, causing relaxations in the bound on $I(\mathsf{x},\mathsf{y})$. We may therefore wish to reduce the effects which the noise of the stochastic projection $p(\mathsf{y}|\mathsf{x})$ has on the specification of the decoder $q(\mathsf{x}|\mathsf{y})$. One way to address this matter is by introducing an additional mapping $p(\mathsf{z}|\mathsf{x},\mathsf{y})$ to the auxiliary variable space, which may be thought of as an additional variational parameter (see Fig. 2). Indeed, even when the channel is noisy, the conditional dependence of the auxiliary variables z on the unperturbed source patterns could result in an accurate detection of the states of the auxiliary variables. Note that the *auxiliary conditional* distribution $p(\mathsf{z}|\mathsf{x},\mathsf{y})$ is defined in a way that does not affect the original noisy channel $p(\mathsf{y}|\mathsf{x})$, as the channel would remain a marginal of the joint distribution of the original sources, codes, and auxiliary variables

$$p(\mathsf{x},\mathsf{y},\mathsf{z}) = \tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})p(\mathsf{z}|\mathsf{x},\mathsf{y}). \tag{5}$$

The role of the auxiliary variables z in this context would be to capture global *features* of the transmitted sources, and use these features for choosing optimal experts for the decoder. Importantly, the auxiliary variables z are *not* transmitted across the channel. Their purpose here is to define a richer family of bounds on $I(\mathsf{x},\mathsf{y})$ which would generalize over objectives with simple constraints on variational decoders (such as linear Gaussians).

From the definition (5) and the chain rule for mutual information (e.g. [10]), we may express $I(\mathsf{y},\mathsf{x})$ as

$$I(\mathsf{y},\mathsf{x}) = I(\{\mathsf{z},\mathsf{y}\},\mathsf{x}) - I(\mathsf{x},\mathsf{z}|\mathsf{y}), \tag{6}$$

where $I(\{\mathsf{z},\mathsf{y}\},\mathsf{x}) \stackrel{\text{def}}{=} H(\mathsf{x}) - H(\mathsf{x}|\mathsf{z},\mathsf{y})$ is the amount of information that the features z and codes y jointly contain about the sources, and $I(\mathsf{x},\mathsf{z}|\mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{z}|\mathsf{y}) - H(\mathsf{z}|\mathsf{x},\mathsf{y})$ is the conditional mutual information. Substituting the definitions into (6), we obtain a general expression of the mutual information $I(\mathsf{x},\mathsf{y})$ as a function of conditional entropies of the sources, codes, and auxiliary variables

$$I(\mathsf{y},\mathsf{x}) = H(\mathsf{x}) + H(\mathsf{z}|\mathsf{x},\mathsf{y}) - H(\mathsf{x}|\mathsf{y},\mathsf{z}) - H(\mathsf{z}|\mathsf{y}). \tag{7}$$

Then by analogy with (4) we obtain

$$I(\mathsf{y},\mathsf{x}) \geq H(\mathsf{x}) + H(\mathsf{z}|\mathsf{x},\mathsf{y}) + \langle \log q(\mathsf{x}|\mathsf{y},\mathsf{z}) \rangle_{p(\mathsf{x},\mathsf{y},\mathsf{z})} + \langle \log q(\mathsf{z}|\mathsf{y}) \rangle_{p(\mathsf{y},\mathsf{z})}. \tag{8}$$

Symbolically, (8) has a form vaguely reminiscent of the objectives optimized by *Information Bottleneck* (IB) methods [15]. However, the similarity is deceptive both conceptually and analytically, which is easy to see by comparing the objectives and the extrema. Additionally, we note that the auxiliary variational method is applicable to significantly more complex channels, provided that the variational distributions are appropriately constrained.

The mapping $p(\mathsf{z}|\mathsf{x},\mathsf{y})$ to the feature space may be constrained so that the averages in (8) are tractable, e.g.

$$p(z_j|\mathsf{x},\mathsf{y}) = p(z_j|\mathsf{x}) \propto \exp\{-(\mathsf{v}_j^T\mathsf{x} + b_j)\}, \tag{9}$$

where z_j is the j^{th} state of a multinomial variable z. Analogously, we may constrain the variational decoders $q(\mathsf{x}|\mathsf{y},\mathsf{z})$ and $q(\mathsf{z}|\mathsf{y})$. In a specific case of a linear Gaussian channel $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}(\mathsf{W}\mathsf{x}, s^2\mathsf{I})$, we may assume $q(\mathsf{x},z|\mathsf{y}) \propto q(\mathsf{x}|\mathsf{y},z)q(z)$ with $q(\mathsf{x}|\mathsf{y},z_j) \sim \mathcal{N}(\mathsf{U}_j\mathsf{y},\mathsf{S}_j)$. Then objective (8) is optimized for the channel encoder, variational decoder, and the auxiliary conditional distributions, which is tractable for the considered parameterization. Effectively, we will still be learning a noisy linear projection, but for a different (mixture-type) variational decoder.

3.1 Learning Representations in the Augmented $\{\mathbf{y}, z\}$-space

Now suppose that the multinomial auxiliary variable z is actually observable at the receiver's end of the channel. Under this assumption, we may consider optimizing an alternative bound $\tilde{I}_H(\mathsf{x},\{\mathsf{y},z\}) \geq I(\mathsf{x},\mathsf{y})$, defined by analogy with

(4). (We will use the notation I_H to indicate that the channel $\mathsf{x} \to \{\mathsf{y}, z\}$ is generally heterogeneous; for example, z may be a generally unknown class label, while $\mathsf{y} \in \mathbb{R}^{|\mathsf{y}|}$ may define a lower-dimensional projection). This leads to a slight simplification of (8), which effectively reduces to

$$\tilde{I}_H(\mathsf{x}, \{\mathsf{y}, z\}) = H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y}, z) \rangle_{\tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})p(z|\mathsf{x})}, \tag{10}$$

where the cross-entropic term is given by

$$\begin{aligned} \langle \log q(\mathsf{x}|\mathsf{y}, z) \rangle_{p(\mathsf{x},\mathsf{y},z)} = &-\frac{1}{2M} \sum_{j=1}^{|z|} \sum_{i=1}^{M} p(z_j|\mathsf{x}^{(i)}) \mathrm{tr} \left\{ \mathsf{S}_j^{-1} \left(\mathsf{d}_j^{(i)} \mathsf{d}_j^{(i)T} + s^2 \mathsf{U}_j \mathsf{U}_j^T \right) \right\} \\ &-\frac{1}{2M} \sum_{j=1}^{|z|} \log |\mathsf{S}_j| \sum_{i=1}^{M} p(z_j|\mathsf{x}^{(i)}). \end{aligned} \tag{11}$$

Here we ignored the irrelevant constants and defined

$$\mathsf{d}_j^{(i)} \stackrel{\text{def}}{=} \mathsf{x}^{(i)} - \mathsf{U}_j \mathsf{W} \mathsf{x}^{(i)} \in \mathbb{R}^{|\mathsf{x}|} \tag{12}$$

to be the distortion between the i^{th} pattern and its reconstruction from a noiseless code at the mean of $q(\mathsf{x}|\mathsf{y}, z_j)$. From (10) and (12) it is easy to see that small values of the distortion terms $\mathsf{d}_j^{(i)}$ lead to improvements in the bound on $I(\mathsf{x}, \{\mathsf{y}, z\})$, which agrees with the intuition that the trained model should favour accurate reconstructions of the source patterns from their compressed representations.

Note that in the communication-theoretic interpretation of the considered heterogeneous channel, the auxiliary variables z will need to be communicated over the channel (*cf* bound (8)). Generally, this comes at a small increase in the communication cost, which in this case is $\sim O(|z|)$. For the model parameterization considered here, this corresponds to sending (or storing) an additional positive integer z, which would effectively index the decoder used at the reconstruction. Generally, the lower-dimensional representations of $\{\mathsf{x}\}$ will include not only the codes $\{\mathsf{y}\}$, but also the auxiliary labels z. Finally, we may note that unless $p(z|\mathsf{x})$ is strongly constrained, the mapping $\mathsf{x} \to z$ will typically tend to be nearly noiseless, as this would decrease $H(z|\mathsf{x})$ and maximize $I(\mathsf{x}, \{\mathsf{y}, z\})$.

4 Demonstrations

Here we demonstrate a few applications of the method to extracting optimal subspaces for the digits dataset. In all cases, it was assumed that $|\mathsf{y}| < |\mathsf{x}|$. We also assumed that $p(\mathsf{x})$ is the empirical distribution.

4.1 Hand-Written Digits: Comparing the Bounds

In the first set of experiments, we compared optimal lower bounds on the mutual information $I(\mathsf{x}, \mathsf{y})$ obtained by maximizing the as-if Gaussian $I_G(\mathsf{x}, \mathsf{y})$ and

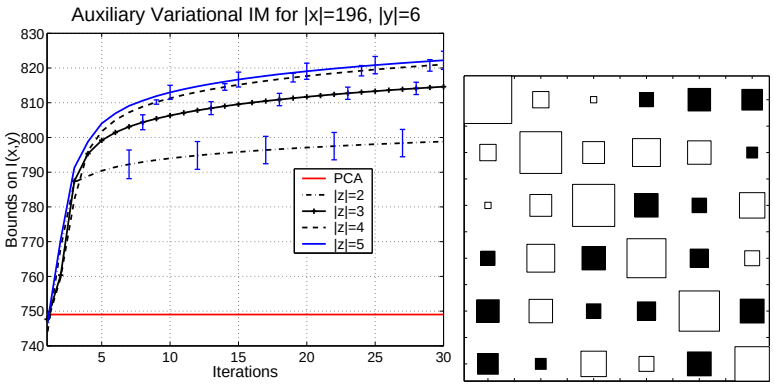


Fig. 3. Variational information maximization for noisy constrained dimensionality reduction. **(a):** *Top curves:* Average values of the variational auxiliary bounds $\tilde{I}(\mathsf{x},\mathsf{y})$, obtained by the IM algorithm started at 10 random model initializations (shown for $|z| = 2,\ldots,5$); *bottom line*: the *as-if* Gaussian $I_G(\mathsf{x},\mathsf{y})$ bound (computed numerically). The results are shown for the digits data with $|\mathsf{x}| = 196$, $|\mathsf{y}| = 6$ for $M = 30$ patterns and $T = 30$ iterations of the IM. **(b):** Hinton diagram for $\mathsf{W}\mathsf{W}_{pca}^T(\mathsf{W}\mathsf{W}_{pca}^T)^T \in \mathbb{R}^{6\times 6}$ for $|z| = 3$, $T = 30$. For orthonormal weights spanning identical subspaces, we would expect to see the identity matrix.

the auxiliary variational $\tilde{I}(\mathsf{x},\mathsf{y})$ objectives for hand-written digits. The dataset contained $M = 30$ gray-scaled instances of 14-by-14 digits 1, 2, and 8 (10 of each class), which were centered and normalized. The goal was to find a noisy projection of the $|\mathsf{x}| = 196$-dimensional training data into a $|\mathsf{y}| = 6$-dimensional space, so that the bounds $I_G(\mathsf{x},\mathsf{y})$ and $\tilde{I}(\mathsf{x},\mathsf{y})$ were maximized. We considered a linear Gaussian channel with an irreducible white noise, which in this case leads to the encoder distribution $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_y\left(\mathsf{W}\mathsf{y}, s^2\mathsf{I}\right)$ with $\mathsf{W} \in \mathbb{R}^{6\times 196}$. Our specific interest was in finding optimal *orthonormal* projections, so the weights were normalized to satisfy $\mathsf{W}\mathsf{W}^T = \mathsf{I}_{|\mathsf{y}|}$ (by considering the parameterization $\mathsf{W} = (\tilde{\mathsf{W}}\tilde{\mathsf{W}}^T)^{-1/2}\tilde{\mathsf{W}}$ with $\tilde{\mathsf{W}} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$). Effectively, this case corresponds to finding the most informative compressed representations of the source vectors for improving communication of the *non-Gaussian* data over a noisy Gaussian channel (by maximizing lower bounds on the channel capacity). Our specific interest here was to find whether we may indeed improve on Linsker's as-if Gaussian bound on the mutual information (with the optima given in this case by the PCA projection) by considering a richer family of auxiliary variational bounds with multi-modal mixture-type decoders.

Figure 3 shows typical changes in the auxiliary variational bound $\tilde{I}(\mathsf{x},\mathsf{y})$ as a function of the IM's iterations T for $|z| \in \{2,\ldots,5\}$ states of the discrete auxiliary variable. (On the plot, we ignored the irrelevant constants $H(\mathsf{x})$ identical for both $\tilde{I}(\mathsf{x},\mathsf{y})$ and $I_G(\mathsf{x},\mathsf{y})$, and interpolated $\tilde{I}(\mathsf{x},\mathsf{y})$ for the consecutive iterations).

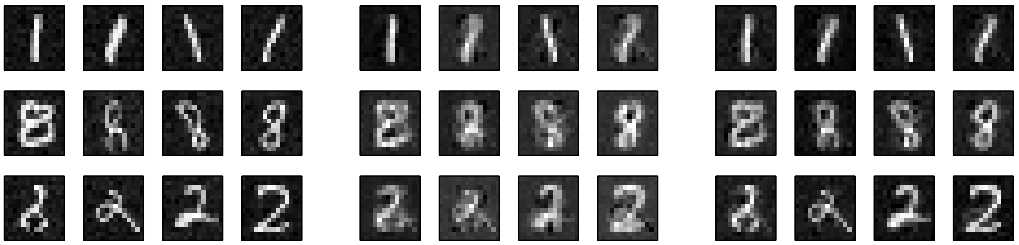
The mappings were parameterized as described in Section 3, with the random initializations of the parameters v_j and b_j around zero, and the initial settings of the variational prior $q(z) = 1/|z|$. The encoder weights W were initialized at 6 normalized principal components $\mathsf{W}_{pca} \in \mathbb{R}^{6\times 196}$ of the sample covariance $\langle \mathsf{x}\mathsf{x}^T \rangle$, and the variance of the channel noise was fixed at $s^2 = 1$. For each choice of the auxiliary space dimension $|z|$, Figure 3 (a) shows the results averaged over 30 random initializations of the IM algorithm. As we see from the plot, the IM learning leads to a consistent improvement in the auxiliary variational bound, which (on average) varies from $\tilde{I}_0(\mathsf{x},\mathsf{y}) \approx 745.7$ to $\tilde{I}_T(\mathsf{x},\mathsf{y}) \approx 822.2$ at $T = 30$ for $|z| = 5$. Small variances in the obtained bounds ($\sigma_T \approx 2.6$ for $T = 30$, $|z| = 5$) indicate a stable increase in the information content independently of the model initializations. From Figure 3 (a) we can also observe a consistent improvement in the average $\tilde{I}(\mathsf{x},\mathsf{y})$ with $|z|$, changing as $\tilde{I}_{10}(\mathsf{x},\mathsf{y}) \approx 793.9$, ≈ 806.3, ≈ 811.2, and ≈ 812.9 for $|z| = 2,\ldots,5$ after $T = 10$ IM's iterations. In comparison, the PCA projection weights W_{pca} result in $I_G(\mathsf{x},\mathsf{y}) \approx 749.0$, which is visibly worse than the auxiliary bound with the optimized parameters, and is just a little better than $\tilde{I}(\mathsf{x},\mathsf{y})$ computed at a random initialization of the variational decoder for $|z| \geq 2$.

Importantly, we stress once again that the auxiliary variables z are not passed through the channel. In the specific case which we considered here, the auxiliary variables were used to define a more powerful family of variational bounds which we used to extract the $\tilde{I}$-optimal orthonormal subspace. The results are encouraging, as they show that for a specific constrained channel distribution we may indeed obtain more accurate lower bounds on the mutual information $I(\mathsf{x},\mathsf{y})$ *without* communicating more data than in the conventional case. Specifically, for Gaussian channels with orthonormal projections to the code space, we do improve on simple *as-if Gaussian* solutions (leading to the PCA projections) by considering optimization of the auxiliary variational bounds (8).

As expected, we may also note that the $\tilde{I}$-optimal *encoder* weights W are in general different from rotations of W_{pca}. This is easy to see by computing $\mathsf{W}\mathsf{W}_{pca}^T(\mathsf{W}\mathsf{W}_{pca}^T)^T$, which in our case is visibly different from the identity matrix (see Fig. 3 (b) for $|\mathsf{y}| = 6$ and $|z| = 3$), which we would have expected to obtain otherwise. This indicates the intuitive result that by allowing a greater flexibility in the choice of the *variational decoder* distributions, the $\tilde{I}(\mathsf{x},\mathsf{y})$-optimal constrained *encoders* become different from the optimal encoders of simpler models (such as PCA under the linear Gaussian assumption).

4.2 Hand-Written Digits: Reconstructions

Additionally, for the problem settings described in Sec. 4.1, we have computed reconstructions of the source patterns $\{\mathsf{x}\}$ from their noisy encoded representations. First, we generated source vectors by adding an isotropic Gaussian noise to the generic patterns (see Fig. 4 (a)), where the variance of the source noise was set as $s_s^2 = 0.5$. Then we computed noisy linear projections $\{\mathsf{y}\}$ of the source vectors by using the I_G- and the $\tilde{I}_H$- optimal encoder weights (in the latter case, we also computed the auxiliary label z by sampling from the learned $p(z|\mathsf{x})$).



(a) (b) (c)

Fig. 4. Reconstructions of the source patterns from encoded representations. **(a):** A subset of the generic patterns used to generate the source vectors; **(b):** the corresponding reconstructions from 6 principal components; **(c):** the corresponding $\tilde{I}_H$-optimal reconstructions at $\langle \mathsf{x} \rangle_{q(\mathsf{x}|\mathsf{y},z)} = \mathsf{U}_z \mathsf{y}$ for the hybrid $\{\mathsf{y}, z\}$ representations ($|\mathsf{y}| = 6$, $|z| = 3$).

This stage corresponds to passing encoded representations over the noisy channels, where the noise variance for the Gaussian part of the channel was fixed at $s^2 = 1$. Finally, we have used the optimal *approximate* decoders to perform the reconstructions from $\{\mathsf{y}\}$ (for I_G-optimal PCA projections) and $\{\mathsf{y}, z\}$ (for $\tilde{I}_H$-optimal hybrid channels).

As we see from Figure 4 (b), (c), by a slight modification of the channel (due to encoding and communicating a multinomial variable z), we may achieve a visible improvement in the reconstruction of the sources by using the $\tilde{I}_H$- optimal projections[4]. The results are shown for $|\mathsf{y}| = 6$, $|z| = 3$ after $T = 3$ iterations, and the reconstructions are computed at the analytical mean of the decoder's component $q(\mathsf{x}|\mathsf{y}, z)$ indexed by the auxiliary variable z. Even though the resulting hybrid channel may be difficult to justify from the communication viewpoint, the results suggest that maximization of the bound on $I(\mathsf{x}, \{\mathsf{y}, z\})$ provides a sensible way to reduce dimensionality of the sources for the purpose of reconstructing inherently noisy non-Gaussian patterns. Importantly, the variational decoder $q(z|\mathsf{x}, \mathsf{y})$ which maximizes $\tilde{I}_H(\mathsf{x}, \{\mathsf{y}, z\})$ makes no recourse to $\tilde{p}(\mathsf{x})$. Therefore, just like in the PCA case, we do not need to store the training instances in order to perform an accurate reconstruction from noisy lower-dimensional projections. We note once again that the weights of the optimal encoder were chosen to satisfy the specific orthonormality constraint (though other kinds of constrained encoders may easily be considered). This contrasts with the exact approaches to training generative models, where encoding constraints may be more difficult to enforce.

[4] Similar reconstructions could be obtained by maximizing the auxiliary bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ *without* communicating z. However, the approximate decoder for this case would be given as $q(\mathsf{x}|\mathsf{y}) = \sum_z q(\mathsf{x}|\mathsf{y}, z) \frac{\langle p(z|\mathsf{x}) p(\mathsf{y}|\mathsf{x}) \rangle_{p(\mathsf{x})}}{\langle p(z|\mathsf{x}) \rangle_{p(\mathsf{x})}}$, which requires knowing $p(\mathsf{x})$.

5 Summary

Here we described an auxiliary variational approach to information maximization, and applied it to linear orthonormal dimensionality reduction in the presence of irreducible Gaussian noise. For this case we showed that the common *as-if* Gaussian [13] approximation of MI is in fact a suboptimal special case of our variational bound, which for isotropic linear Gaussian channels leads to the PCA solution. Importantly, this means that by using linear Gaussian variational decoders under the considered Gaussian channel, maximization of the generic lower bound (4) on MI cannot improve on the PCA projections. The situation changes if we consider a richer family of *variational auxiliary* lower bounds on $I(\mathsf{x}, \mathsf{y})$ under the same encoding constraints. In particular, we showed that in the cases when the source distribution was non-Gaussian, we could significantly improve on the PCA projections by considering multi-modal variational decoders. This confirms the conceptually simple idea that by allowing a greater flexibility in the choice of variational decoders, we may get significant improvements over simple bounds at a limited increase in the computational cost. This result is also interesting from the communication-theoretic perspective, as it demonstrates a simple and computationally efficient way to produce better bounds on the capacity of communication channels without altering channel properties (e.g. without communicating more data across the channels). Finally, we discussed a simple information-theoretic approach to constrained dimensionality reduction for hybrid representations $\mathsf{x} \to \{\mathsf{y}, z\}$, which may significantly improve reconstructions of the sources $\{\mathsf{x}\}$ from their lower-dimensional representations $\{\mathsf{y}\}$ at a small increase in the transmission cost (given by $|z|$).

It is potentially interesting to compare the variational information-maximizing framework with other approaches applicable to learning unknown under-complete representations of the data (such as generative models[5] and autoencoders). As we pointed out, there are important conceptual differences in the way we parameterize and train encoder and generative models. Specifically, by imposing explicit constraint on the mapping to the space of representations, our method is applicable for *constrained stochastic* dimensionality reduction. This may be particularly useful in engineering and neural systems, where such constraints may be physically or biologically motivated. Despite the important differences, it is interesting to note that the special case of the auxiliary variational bound on $I(\mathsf{x}, \mathsf{y})$ for a Gaussian channel and a multinomial auxiliary space $\{z\}$ has an interesting link to likelihood maximization for a mixture of factor-analysis-type models with the *uniform*, rather than Gaussian, factor distribution [2] (*cf* [14]).

It is also interesting to compare our framework with self-supervised training in semi-parametric models. The most common application of self-supervised models is dimensionality reduction in autoencoders $\mathsf{x} \to \mathsf{y} \to \tilde{\mathsf{x}}$, where $\mathsf{x}^{(m)} = \tilde{\mathsf{x}}^{(m)}$ for all patterns m. Typically, it is presumed that $\mathsf{y} = \mathsf{f}(\mathsf{x})$, and the models are

[5] It is well known that in a few special cases (e.g. for square ICA models) mutual information- and likelihood-maximization may lead to the same extrema [9]. However, little is understood about how the frameworks may relate in general.

trained by minimizing a loss function (such as the squared loss). It is clear that for noiseless encoders, our bound (4) gives $const + \sum_m \log q\left(\mathsf{x}^{(m)}|\mathsf{y} = \mathsf{f}(\mathsf{x}^{(m)})\right)$, which has the interpretation of an autoencoder whose loss function is determined by q. Thus a squared loss function can be interpreted as an assumption that the data x can be reconstructed from *noiseless* codes y with Gaussian fluctuations. However, in some sense, the natural loss function (from the MI viewpoint) would not be the squared loss, but that which corresponds to the Bayesian decoder $q(\mathsf{x}|\mathsf{y}) = p(\mathsf{x}|\mathsf{y})$, and more powerful models should strive to approximate this. Indeed, this is also the role of the auxiliary variables – effectively to make a loss function which is closer to the Bayes optimum. What is also interesting about our framework is that it holds in the case that the codes are stochastic, for which the autoencoder framework is more clumsy. Indeed, it also works when we have a (non-delta mixture) distribution $p(\mathsf{x})$, i.e. the method merges many interesting models in one framework.

References

1. Agakov, F. V. and Barber, D (2004). Variational Information Maximization for Population Coding. In *ICONIP*.
2. Agakov, F. V. Variational Information Maximization in Stochastic Environments. Submitted.
3. Arimoto, S. (1972). Algorithm for computing the capacity of arbitrary discrete memoryless channels. *IT*, 18.
4. Barber, D. and Agakov, F. V. (2003). The IM Algorithm: A Variational Approach to Information Maximization. In *NIPS*.
5. Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.
6. Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IT*, 18.
7. Bishop, C. M. and Svensen, M. and Williams, C. W. I. (1998). GTM: The Generative Topographic Mapping. *Neural Computation*, 10(1):215–234.
8. Brunel, N. and Nadal, J.-P. (1998). Mutual Information, Fisher Information and Population Coding. *Neural Computation*, 10:1731–1757.
9. Cardoso, J. F. (1997). Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4.
10. Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.
11. Jacobs, R. A. and Jordan, M. I. and Nowlan, S. J. and Hinton, G. E. (1991). Adaptive Mixtures of Local Experts. *Neural Computation*, 3.
12. Linsker, R. (1989). An Application of the Principle of Maximum Information to Linear Systems. In *NIPS*.
13. Linsker, R. (1993). Deriving Receptive Fields Using an Optimal Encoding Criterion. In *NIPS*.
14. Tipping, M. E. and Bishop, C. M. (1999a). Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2): 443–482.
15. Tishby, N. and Pereira, F. and Bialek, W. (1999) The information bottleneck method. In *Proc. of the 37-th AACCCC*.
16. Williams, C. K. I. and Agakov, F. V. (2002). Products of Gaussians and Probabilistic Minor Component Analysis. *Neural Computation*, 14(5).