

# Auxiliary Variational Information Maximization for Dimensionality Reduction

Felix Agakov<sup>1</sup> and David Barber<sup>2</sup>

<sup>1</sup> University of Edinburgh, 5 Forrest Hill, EH1 2QL Edinburgh, UK  
felixa@inf.ed.ac.uk, www.anc.ed.ac.uk

<sup>2</sup> 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 compute, and most previous studies redefine the criterion in forms of approximations. Recently we described properties of a simple lower bound on MI [2], and discussed its links to some of the popular dimensionality reduction techniques. Here we introduce a richer family of the *auxiliary variational* bounds on MI, which generalize our previous approximations. Our specific focus then is on applying the bound to extracting informative lower-dimensional orthonormal projections in the presence of irreducible Gaussian noise. We show that our method produces significantly tighter bounds than the *as-if Gaussian* approximations [7] of MI. We also show that learning projections to multinomial auxiliary spaces may facilitate reconstructions of the sources from noisy lower-dimensional representations.

## 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 is not lost. If it is not known a priori which coordinates of  $x$  may be relevant for a specific tasks, 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

$$I(x, y) \equiv H(x) - H(x|y), \quad (1)$$

which indicates the decrease of uncertainty in  $x$  due to the knowledge of  $y$ . Here  $H(x) \equiv -\langle \log p(x) \rangle_{p(x)}$  and  $H(x|y) \equiv -\langle \log p(x|y) \rangle_{p(x,y)}$  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  $p(y|x)$ . However, it is easy to see that if the dimension  $|y|$  ( $|y| < |x|$ ) is still large, the exact evaluation of  $I(x, y)$  is in general computationally intractable. The key difficulty lies in the computation of the conditional entropy  $H(x|y)$ , which is tractable only in a few special cases. Typically, the standard techniques assume that  $p(x, y)$

is jointly Gaussian, the output spaces are very low-D, or the channels are deterministic and invertible [6], [3]. Unfortunately, these assumptions may be too restrictive for many practical applications. Other methods suggest alternative objective functions (e.g. approximations based on the *Fisher Information* [4]), which, however, do not retain proper bounds on  $I(\mathbf{x}, \mathbf{y})$  and may often lead to numerical instabilities when applied to dimensionality reduction [1].

### 1.1 Linsker’s *as-if Gaussian* approximation

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

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

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

$$2I_G(\mathbf{x}, \mathbf{y}) = \log |\boldsymbol{\Sigma}_{xx}| - \log |\boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{xy}^T|. \quad (3)$$

Here  $\boldsymbol{\Sigma}_{xx}$ ,  $\boldsymbol{\Sigma}_{xy}$ , and  $\boldsymbol{\Sigma}_{yy}$  are the partitions of  $\boldsymbol{\Sigma}$ . Clearly, the objective (3) is a function of the encoder parameters. After training, the encoder may be used for producing lower-dimensional representations  $\mathbf{y}$  for a given source  $\mathbf{x}$ .

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

A simple lower bound on the mutual information  $I(\mathbf{x}, \mathbf{y})$  follows from non-negativity of the Kullback-Leibler divergence  $KL(p(\mathbf{x}|\mathbf{y})||q(\mathbf{x}|\mathbf{y}))$  between the exact posterior  $p(\mathbf{x}|\mathbf{y})$  and its variational approximation  $q(\mathbf{x}|\mathbf{y})$ , leading to

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

Here  $q(\mathbf{x}|\mathbf{y})$  is an arbitrary distribution saturating the bound for  $q(\mathbf{x}|\mathbf{y}) \equiv p(\mathbf{x}|\mathbf{y})$ . The objective (4) explicitly includes<sup>1</sup> both the encoder  $p(\mathbf{y}|\mathbf{x})$  (distribution of the lower-D representations for a given source) and decoder  $q(\mathbf{x}|\mathbf{y})$  (reconstruction of the source from a given compressed representation). The flexibility of the choice of the decoder  $q(\mathbf{x}|\mathbf{y})$  makes (4) particularly computationally convenient. Typically, we will be interested in constraining  $q(\mathbf{x}|\mathbf{y})$  to lie in a tractable family.

It is easy to show that by constraining the decoder as  $q(\mathbf{x}|\mathbf{y}) \sim \mathcal{N}(\mathbf{U}\mathbf{y}, \boldsymbol{\Sigma})$ , optimization of the bound (4) reduces to maximization of Linsker’s *as-if Gaussian* approximation (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  $q(\mathbf{x}|\mathbf{y})$  is a linear Gaussian. Moreover, if for this case  $p(\mathbf{y}|\mathbf{x}) \sim \mathcal{N}(\mathbf{W}\mathbf{x}, s^2\mathbf{I})$ ,

<sup>1</sup> The bound (4) differs from the standard Blahut-Arimoto algorithms (e.g. [5]) in the sense that  $q(\mathbf{x}|\mathbf{y})$  is constrained to be tractable (i.e.  $q(\mathbf{x}|\mathbf{y}) \neq p(\mathbf{x}|\mathbf{y})$  in general). The variational *IM* algorithm [2] optimizes (4) for parameters of  $q(\mathbf{x}|\mathbf{y})$  and  $p(\mathbf{y}|\mathbf{x})$ .

and  $p(\mathbf{x})$  is the empirical distribution of the training patterns, it is easy to show that the left singular vectors of the optimal projection weights  $W^T$  correspond to the  $|\mathbf{y}|$ -PCA solution on the sample covariance  $\langle \mathbf{x}\mathbf{x}^T \rangle$ .

### 3 An Auxiliary Variational Bound

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

Let  $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x}, \mathbf{y})p(\mathbf{z}|\mathbf{x}, \mathbf{y})$  define a general joint distribution over the sources  $\mathbf{x}$ , the encodings  $\mathbf{y}$ , and some auxiliary (“feature”) variables  $\mathbf{z}$ . Then from the chain rule for mutual information (e.g. [5]) we get

$$I(\mathbf{y}, \mathbf{x}) = I(\{\mathbf{z}, \mathbf{y}\}, \mathbf{x}) - I(\mathbf{x}, \mathbf{z}|\mathbf{y}), \quad (5)$$

where  $I(\{\mathbf{z}, \mathbf{y}\}, \mathbf{x}) \equiv H(\mathbf{x}) - H(\mathbf{x}|\mathbf{z}, \mathbf{y})$  is the amount of information that the features  $\mathbf{z}$  and codes  $\mathbf{y}$  jointly contain about the sources, and  $I(\mathbf{x}, \mathbf{z}|\mathbf{y}) \equiv H(\mathbf{z}|\mathbf{y}) - H(\mathbf{z}|\mathbf{x}, \mathbf{y})$  is the conditional mutual information. By analogy with (4), we get

$$I(\mathbf{y}, \mathbf{x}) \geq H(\mathbf{x}) + H(\mathbf{z}|\mathbf{x}, \mathbf{y}) + \langle \log q(\mathbf{x}|\mathbf{y}, \mathbf{z}) \rangle_{p(\mathbf{x}, \mathbf{y}, \mathbf{z})} + \langle \log q(\mathbf{z}|\mathbf{y}) \rangle_{p(\mathbf{y}, \mathbf{z})}. \quad (6)$$

The mapping  $p(\mathbf{z}|\mathbf{x}, \mathbf{y})$  to the feature space may be constrained so that the averages in (6) are tractable, e.g.

$$p(z_j|\mathbf{x}, \mathbf{y}) = p(z_j|\mathbf{x}) \propto \exp\{-(\mathbf{v}_j^T \mathbf{x} + b_j)\}, \quad (7)$$

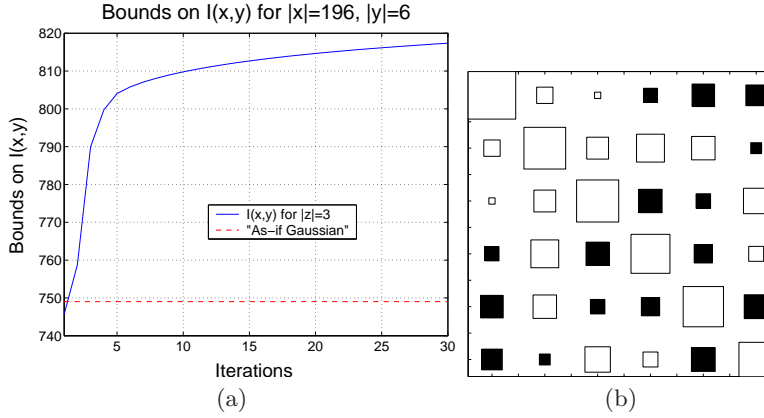
where  $z_j$  is the  $j^{\text{th}}$  state of a multinomial variable  $z$ . Analogously, we may constrain the variational decoders  $q(\mathbf{x}|\mathbf{y}, \mathbf{z})$  and  $q(\mathbf{z}|\mathbf{y})$ . In a specific case of a linear Gaussian channel  $p(\mathbf{y}|\mathbf{x}) \sim \mathcal{N}(\mathbf{W}\mathbf{x}, s^2\mathbf{I})$ , we may assume  $q(\mathbf{x}, \mathbf{z}|\mathbf{y}) \propto q(\mathbf{x}|\mathbf{y}, \mathbf{z})q(\mathbf{z}|\mathbf{y})$  with  $q(\mathbf{x}|\mathbf{y}, \mathbf{z}_j) \sim \mathcal{N}(\mathbf{U}_j\mathbf{y}, \mathbf{S}_j)$ . Then objective (6) 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<sup>3</sup>.

Alternatively, in practice we may consider optimizing a bound  $\tilde{I}_H(\mathbf{x}, \{\mathbf{y}, \mathbf{z}\}) \geq \tilde{I}(\mathbf{x}, \mathbf{y})$ , defined by analogy with (4). It will lead to a slight simplification of (6). However, in this case the auxiliary variables  $\mathbf{z}$  will need to be communicated over the channel. In other words, the compressed representations of  $\{\mathbf{x}\}$  will include not only  $\{\mathbf{y}\}$ , but also the auxiliary variables  $\{\mathbf{z}\}$ .

### 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  $|\mathbf{y}| < |\mathbf{x}|$ . We also assumed that  $p(\mathbf{x})$  is the empirical distribution.

<sup>3</sup> For some cases, (6) has a form related to the objectives optimized by *Information Bottleneck* methods [8]. However, (6) is more general, as it is applicable to significantly more complex channels. Moreover, there are fundamental conceptual differences in the way the IB objective is defined.

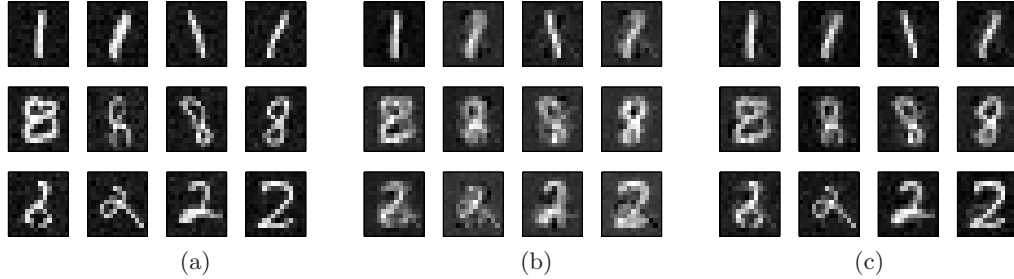


**Fig. 1. (a):** *Solid line:* a typical learning curve for the variational auxiliary  $\tilde{I}(x, y)$  bound on the mutual information (shown for  $|z| = 3$ ); *dashed line:* the *as-if* Gaussian  $I_G(x, y)$  bound (computed numerically). The results are shown for the digits data with  $|x| = 196$ ,  $|y| = 6$  for  $M = 30$  patterns and  $T = 30$  iterations of the IM. **(b):** Hinton diagram for  $WW_{pca}^T(WW_{pca}^T)^T \in \mathbb{R}^{6 \times 6}$ . For orthonormal weights spanning identical subspaces, we would expect the identity matrix.

#### 4.1 Hand-Written Digits: Comparing the Bounds

In the first set of experiments, we compared optimal lower bounds on the mutual information  $I(x, y)$  obtained by maximizing the *as-if* Gaussian  $I_G(x, y)$  and the auxiliary variational  $\tilde{I}(x, y)$  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 an orthogonal projection of the  $|x| = 196$ -dimensional training data into a  $|y| = 6$ -dimensional training space, so that the bounds  $I_G(x, y)$  and  $\tilde{I}(x, y)$  were maximized. As before, we considered a linear Gaussian channel with an irreducible white noise, which in this case leads to the encoder distribution  $p(y|x) \sim \mathcal{N}_y(Wy, s^2 \mathbf{1})$  with  $W \in \mathbb{R}^{6 \times 196}$ . Our interest was in finding optimal orthogonal projections, so the weights were normalized to satisfy  $WW^T = \mathbf{I}_{|y|}$  (by considering the parameterization  $W = (\tilde{W}\tilde{W}^T)^{-1/2}\tilde{W}$  with  $\tilde{W} \in \mathbb{R}^{|y| \times |x|}$ ). 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 the *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 1(a) shows a typical change in the auxiliary variational bound  $\tilde{I}(x, y)$  as a function of the IM's iterations  $T$  for  $|z| = 3$  states of the discrete auxiliary variable. (On the plot, we ignored the irrelevant constants  $H(x)$  identical for both  $\tilde{I}(x, y)$  and  $I_G(x, y)$ , and interpolated  $\tilde{I}(x, y)$  for the consecutive iterations). The mappings were parameterized as described in Sec. 3, with the random ini-



**Fig. 2.** Reconstructions of the source patterns from encoded representations. **(a):** Some of the generic patterns used to generate the source vectors; **(b):** the corresponding reconstructions from 6 principal components; **(c):** the corresponding reconstructions at  $\langle x \rangle_{q(x|y,z)}$  for the hybrid  $\{y, z\}$  representations ( $|y| = 6$ ,  $|z| = 3$ ).

tializations of the parameters  $\mathbf{v}_j$  and  $\mathbf{b}_j$  around zero, and the initial settings of the variational prior  $q(z) = 1/3$ . The encoder weights  $\mathbf{W}$  were initialized at 6 normalized principal components  $\mathbf{W}_{pca} \in \mathbb{R}^{6 \times 196}$  of the sample covariance  $\langle \mathbf{x}\mathbf{x}^T \rangle$ , and the variance of the channel noise was fixed at  $s^2 = 1$ . As we see from Fig. 1 (a), learning leads to a consistent improvement in the auxiliary variational bound, which varies from  $\tilde{I}_0(\mathbf{x}, \mathbf{y}) \approx 745.7$  to  $\tilde{I}_T(\mathbf{x}, \mathbf{y}) \approx 817.4$  at  $T = 30$ . In comparison, the PCA projection weights  $\mathbf{W}_{pca}$  result in  $I_G(\mathbf{x}, \mathbf{y}) \approx 749.0$ , which is visibly worse than the auxiliary bound with the optimized parameters, and is just a little better than  $\tilde{I}(\mathbf{x}, \mathbf{y})$  computed at a random initialization. Importantly, the auxiliary variables  $z$ 's are not passed through the channel; in the specific case which we considered here they were used to define a more powerful family of variational decoders which we used to extract the optimal subspace. The results are encouraging, as they show that for a given Gaussian channel we may indeed obtain tighter bounds on the mutual information (compared with PCA) *without* communicating more data than in the conventional case.

Finally, we note that, as expected, the  $\tilde{I}$ -optimal encoder weights  $\mathbf{W}$  are in general different from rotations of  $\mathbf{W}_{pca}$ . This is easy to see by computing  $\mathbf{W}\mathbf{W}_{pca}^T(\mathbf{W}\mathbf{W}_{pca}^T)^T$ , which in our case is visibly different from the identity matrix (see Fig. 1 (b)), which we would have expected to obtain otherwise.

## 4.2 Hand-Written Digits: Reconstructions

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

As we see from Fig. 2 (b), (c), by a slight modification of the channel (due to encoding and communicating a multinomial variable  $z$ ), we may achieve a significant improvement in the reconstruction of the sources by using the  $\tilde{I}_H$ -optimal projections<sup>4</sup>. Even though the resulting hybrid channel may be difficult to justify from the communication viewpoint, the results suggest that maximization of the bound on  $I(x, \{y, z\})$  provides a sensible way to reduce dimensionality of the sources for the purpose of reconstructing inherently noisy non-Gaussian patterns. Importantly, the variational decoder  $q(z|x, y)$  which maximizes  $\tilde{I}_H(x, \{y, z\})$  makes no recourse to  $p(x)$ . 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.

## 5 Summary

We described an auxiliary variational approach to information maximization, and applied it to dimensionality reduction. We showed that the popular *as-if* Gaussian approximation of the mutual information is in fact a special case of the variational bound, which for isotropic linear Gaussian channels leads to the PCA solution. We also showed that in the cases when the source distribution is non-Gaussian, we may significantly improve on the PCA projections by considering multi-modal variational decoders. Finally, we pointed out a practical information-theoretic approach to dimensionality reduction with hybrid representations  $\{y, z\}$ , which may significantly improve reconstructions of the sources from their lower-dimensional representations.

## References

1. Agakov, F. V. and Barber, D. (2004). Variational Information Maximization for Population Coding. In *ICONIP*.
2. Barber, D. and Agakov, F. V. (2003). The IM Algorithm: A Variational Approach to Information Maximization. In *NIPS*.
3. Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.
4. Brunel, N. and Nadal, J.-P. (1998). Mutual Information, Fisher Information and Population Coding. *Neural Computation*, 10:1731–1757.
5. Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.
6. Linsker, R. (1989). An Application of the Principle of Maximum Information to Linear Systems. In *NIPS*.
7. Linsker, R. (1993). Deriving Receptive Fields Using an Optimal Encoding Criterion. In *NIPS*.
8. Tishby, N. and Pereira, F. and Bialek, W. (1999) The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*.

---

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