Variational Information Maximization in Gaussian Channels

Felix V. Agakov [a] David Barber [b]

IDIAP–RR 04-88

April 2004

SUBMITTED FOR PUBLICATION

[a] School of Informatics, University of Edinburgh, EH1 2QL, UK

[b] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**

Rue du Simplon 4 P.O. Box 592 1920 Martigny − Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-88

Variational Information Maximization in Gaussian Channels

Felix V. Agakov David Barber

April 2004

submitted for publication

Abstract. Recently, we introduced a simple variational bound on mutual information, that resolves some of the difficulties in the application of information theory to machine learning. Here we study a specific application to Gaussian channels. It is well known that PCA may be viewed as the solution to maximizing information transmission between a high dimensional vector x and its low dimensional representation y. However, such results are based on assumptions of Gaussianity of the sources x. In this paper, we show how our mutual information bound, when applied to this arena, gives PCA solutions, without the need for the Gaussian assumption. Furthermore, it naturally generalizes to providing an objective function for Kernel PCA, enabling the principled selection of kernel parameters.

1 Introduction

Maximization of information transmission in noisy channels is a common problem, ranging from the construction of good error-correcting codes [12] and feature extraction [17] to neural sensory processing [11], [6].

The key idea of information maximization is to choose a mapping from source variables (inputs) x to response variables (outputs) y such that the outputs contain as much information about which of the inputs was transmitted as possible. In a stochastic context, we have a source distribution $p(\mathsf{x})$, and a mapping $p(\mathsf{y}|\mathsf{x})$. The general aim will be to set any adjustable parameters of $p(\mathsf{y}|\mathsf{x})$ in order to maximize information transfer. The principal measure of information transfer in this context is the mutual information defined as

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

Equivalently, we may write $I(\mathsf{x},\mathsf{y}) \equiv H(\mathsf{y}) - H(\mathsf{y}|\mathsf{x})$. Here $H(\mathsf{y}) \equiv -\langle \log p(\mathsf{y})\rangle_{p(\mathsf{y})}$ and $H(\mathsf{y}|\mathsf{x}) \equiv -\langle \log p(\mathsf{y}|\mathsf{x})\rangle_{p(\mathsf{x},\mathsf{y})}$ are marginal and conditional entropies respectively, and angled brackets represent averages. The objective (1) is maximized with respect to parameters of the encoder $p(\mathsf{y}|\mathsf{x})$.

Our specific interest in this paper is when we have a finite set of training points $\{\mathsf{x}_m | m = 1,\ldots,M\}$, which forms the empirical distribution $p(\mathsf{x}) = (1/M)\sum_{m=1}^{M}\delta(\mathsf{x}-\mathsf{x}_m)$. Our aim will be to form constrained representations $p(\mathsf{y}|\mathsf{x})$ that maximize $I(\mathsf{x},\mathsf{y})$. For a continuous encoder $p(\mathsf{y}|\mathsf{x})$, the theoretically optimal unconstrained decoder is given by Bayes' rule:

$$p(\mathsf{x}|\mathsf{y}) = \sum_{k=1}^{M}\delta(\mathsf{x}-\mathsf{x}_k)\omega_k(\mathsf{y}), \quad \omega_k(\mathsf{y}) = \frac{p(\mathsf{y}|\mathsf{x}_k)}{\sum_{m=1}^{M} p(\mathsf{y}|\mathsf{x}_m)},$$

which is a mixture of Dirac delta functions. This reduces to a single delta peak only for the case when stochastic images $\mathcal{I}$ of distinct source vectors $\{\mathsf{x}\}$ are non-intersecting under the encoder $p(\mathsf{y}|\mathsf{x})$, i.e. $\forall i,j \in \{1,\ldots,M\}.i \neq j.\mathcal{I}(\mathsf{x}_i)\cap\mathcal{I}(\mathsf{x}_j) = \emptyset$. Formally, this is never the case if $\forall \mathsf{y}.\forall \mathsf{x}.p(\mathsf{y}|\mathsf{x}) \neq 0$, which leads to a mixture form of $p(\mathsf{x}|\mathsf{y})$.

Despite the conceptual simplicity of the statement, computation of the mutual information is generally intractable for all but special cases. For large-scale systems the key difficulty lies in the computation of the conditional entropy $H(\mathsf{x}|\mathsf{y})$ of the mixture distribution $p(\mathsf{x}|\mathsf{y})$, which is in general NP hard.

Standard approaches address the problem of optimizing (1) by assuming that $p(\mathsf{x},\mathsf{y})$ is jointly Gaussian [10], the output spaces are very low-dimensional [11], or the channels are deterministic and invertible [2]. Other popular methods suggest alternative objective functions (e.g. the Fisher information criterion [4]), which, however, do not preserve proper bounds on $I(\mathsf{x},\mathsf{y})$.

1.1 Variational Lower Bound on Mutual Information

In order to derive a simple lower bound on the mutual information, we consider the Kullback-Leibler divergence $KL(p(\mathsf{x}|\mathsf{y})||q(\mathsf{x}|\mathsf{y}))$ between the posterior $p(\mathsf{x}|\mathsf{y})$ and its variational approximation $q(\mathsf{x}|\mathsf{y})$. Non-negativity of the divergence gives

$$\langle \log p(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{x}|\mathsf{y})} - \langle \log q(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{x}|\mathsf{y})} \geq 0 \;\Rightarrow\; \underbrace{\langle \log p(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{x}|\mathsf{y})p(\mathsf{y})}}_{-H(\mathsf{x}|\mathsf{y})} \geq \langle \log q(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{x}|\mathsf{y})p(\mathsf{y})}\,. \tag{2}$$

This leads to

$$I(\mathsf{x},\mathsf{y}) \geq H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{x},\mathsf{y})} \overset{\text{def}}{=} \tilde{I}(\mathsf{x},\mathsf{y}), \tag{3}$$

where $q(\mathsf{x}|\mathsf{y})$ is an arbitrary variational distribution which saturates the bound for $q(\mathsf{x}|\mathsf{y}) \equiv p(\mathsf{x}|\mathsf{y})$. Note that the objective (3) explicitly includes both the encoder $p(\mathsf{y}|\mathsf{x})$ and decoder $q(\mathsf{x}|\mathsf{y})$.

Other well known lower bounds on the mutual information could be considered [7]. However, our current experience suggests that for certain choices of the decoder $q(\mathsf{x}|\mathsf{y})$ the variational bound considered above is particularly computationally convenient. Moreover, since (3) is based on the KL divergence between the true and the approximating posteriors, it is equivalent to a moment matching approximation of $p(\mathsf{x}|\mathsf{y})$ by $q(\mathsf{x}|\mathsf{y})$. This fact is beneficial in terms of decoding, since the more successful decoding algorithms approximate the mean of the posterior $p(\mathsf{x}|\mathsf{y})$ [13], whilst standard mode matching approaches (such as mean-field theory) typically get trapped in the one of many sub-optimal local minima.

Recently in [1] we discussed several applications of the bound (3) and outlined its basic properties. The principal objective of this paper is to investigate properties of the bound for the case when the decoder $q(\mathsf{x}|\mathsf{y})$ is Gaussian. We will also assume that the encoder $p(\mathsf{y}|\mathsf{x})$ is Gaussian, although this assumption is less critical to the tractable optimization of the objective.

1.2 Gaussian Decoders

In this paper we focus on a simple assumption that the approximate decoder $q(\mathsf{x}|\mathsf{y})$ is a Gaussian. Although the optimal decoder in the considered channels should be a mixture distribution $q(\mathsf{x}|\mathsf{y}) = p(\mathsf{x}|\mathsf{y})$, one may hope that Gaussian decoders $q(\mathsf{x}|\mathsf{y})$ perform well if the codes are not strongly overlapping. We show that optimal Gaussian decoders have a strong relation to popular dimensionality reduction techniques, inducing PCA and kernel PCA solutions as special cases.

In the following sections we will consider increasingly complex encoders and decoders. Initially, however, we demonstrate the application of the bound to the simple case of a linear Gaussian encoder and decoder.

2 Linear Gaussian Decoder: $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_y(\mathsf{Wx}, s^2\mathsf{I})$, $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_x(\mathsf{Uy}, \sigma^2\mathsf{I})$

Let the encoder and decoder be given by $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_y(\mathsf{Wx}, s^2\mathsf{I})$ and $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_x(\mathsf{Uy}, \sigma^2\mathsf{I})$ respectively. Our goal is to learn optimal settings of $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}| \times |\mathsf{x}|}$ and $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}| \times |\mathsf{y}|}$ (for fixed σ^2 and s^2) by maximizing the variational bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ on the mutual information, which in this case is expressed as

$$\tilde{I}(\mathsf{x}, \mathsf{y}) \propto 2\mathrm{tr}\left\{\mathsf{UWS}\right\} - \mathrm{tr}\left\{\mathsf{U}\boldsymbol{\Sigma}\mathsf{U}^T\right\} + c. \tag{4}$$

Here c is a constant, $\mathsf{S} = \langle \mathsf{x}\mathsf{x}^T \rangle = \sum_m \mathsf{x}_m(\mathsf{x}_m)^T/M$ is the sample covariance of the zero-mean data, and

$$\boldsymbol{\Sigma} = \mathsf{I}s^2 + \mathsf{WSW}^T \in \mathbb{R}^{|\mathsf{y}| \times |\mathsf{y}|} \tag{5}$$

is the covariance of the distribution of the responses $p(\mathsf{y})$. In the following we assume that the weights W, U and the sample covariance S are non-singular. Note that we make no assumptions about the distribution of the sources $p(\mathsf{x})$. Unsurprisingly, this objective is closely related to the least squares reconstruction error in a linear autoencoder. What is particularly interesting in this context is that it provides a lower bound on the mutual information.

Nature of optimal solutions

Unconstrained optimization of (4) for the encoder's weights W leads to the extremum condition

$$\mathsf{U}^T\mathsf{S} = \mathsf{U}^T\mathsf{UWS}. \tag{6}$$

By assuming that y is a compressed representation of the source x (i.e. $|\mathsf{x}| > |\mathsf{y}|$), we obtain $\mathsf{W} = (\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T$. This transforms the objective function (4) into

$$\tilde{I}(\mathsf{x},\mathsf{y}) = \mathrm{tr}\left\{\mathsf{UWSW}^T\mathsf{U}^T\right\} - s^2\mathrm{tr}\left\{\mathsf{UU}^T\right\} = \mathrm{tr}\left\{\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{S}\right\} - s^2\mathrm{tr}\left\{\mathsf{UU}^T\right\}, \tag{7}$$

where we ignored the irrelevant constant c.

Let $\mathsf{U} = \mathsf{VLR}^T$ be the singular value decomposition of the decoder weights U where, by definition, $\mathsf{L} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is diagonal and $\mathsf{V}^T\mathsf{V} = \mathsf{R}^T\mathsf{R} = \mathsf{RR}^T = \mathsf{I}_{|\mathsf{y}|}$. From (6) it is clear that $\mathsf{W} = \mathsf{RL}^{-1}\mathsf{V}^T$, i.e. $\mathsf{WU} = \mathsf{I}_{|\mathsf{y}|}$. Substitution into (7) results in

$$\tilde{I}(\mathsf{x},\mathsf{y}) = \mathrm{tr}\left\{\mathsf{V}^T\mathsf{SV}\right\} - s^2\mathrm{tr}\left\{\mathsf{L}^2\right\}. \tag{8}$$

Optimizing (8) for V under the orthonormality constraints on V, we readily obtain the PCA solution (and its rigid rotations). In the limit $s^2 \to 0$, the solutions are invariant with respect to the scalings L.

For noisy channels $s^2 > 0$, maximization of (8) leads to divergence of the Frobenius norm $\|\mathsf{W}\|_F$, so that the contribution of the finite isotropic noise in (5) is negligible. One way to handle the problem of divergence for $s^2 > 0$ is by constraining the singular values of U. In this case optimal weights U correspond to principal eigenvectors of S. Note that solutions are invariant with respect to complimentary rotations of W and U.

This demonstrates the simple result that PCA provides a lower bound on the mutual information between x and y. Again, we stress that this conclusion is reached without the need for a Gaussian assumption about the source distribution. In order to go beyond the PCA solution, more complex encoders/decoders are required. In the following section we consider the effect of increasing the complexity of the encoder, whilst still using a simple linear decoder as above.

3 Nonlinear Gaussian Channels: $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_y(\mathsf{W}\boldsymbol{\phi}(\mathsf{x}), \boldsymbol{\Sigma}_y)$, $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_x(\mathsf{Uy}, \boldsymbol{\Sigma}_x)$

Here we consider the case of a non-linear Gaussian channel with the encoder $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_y(\mathsf{W}\boldsymbol{\phi}(\mathsf{x}), \boldsymbol{\Sigma}_y)$ and decoder $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_x(\mathsf{Uy}, \boldsymbol{\Sigma}_x)$. Note that for all the data points $\{\mathsf{x}_m|m = 1,\ldots,M\}$, the set of encodings $\{\mathsf{y}_m\}$ is given by a noisy linear projection from the (potentially high-dimensional) feature space $\{\boldsymbol{\phi}(\mathsf{x}_m)\}$. In what follows we assume that $|\boldsymbol{\phi}| > M$, i.e. dimensionality of the feature space exceeds the number of training points.

It is easy to see that for the considered case the lower bound on the mutual information $I(\mathsf{x},\mathsf{y})$ is given by

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_x^{-1}\mathsf{S}\right\} + \mathrm{tr}\left\{\boldsymbol{\Sigma}_x^{-1}\mathsf{UW}\langle\boldsymbol{\phi}(\mathsf{x})\mathsf{x}^T\rangle\right\} - \frac{1}{2}\mathrm{tr}\left\{\mathsf{U}^T\boldsymbol{\Sigma}_x^{-1}\mathsf{U}\left(\boldsymbol{\Sigma}_y + \mathsf{W}\langle\boldsymbol{\phi}(\mathsf{x})\boldsymbol{\phi}(\mathsf{x})^T\rangle\mathsf{W}^T\right)\right\}$$

where $\mathsf{S} = \langle\mathsf{xx}^T\rangle = \sum_m \mathsf{x}_m(\mathsf{x}_m)^T/M$ is the sample covariance of the zero-mean data, and the averages are computed with respect to the empirical distribution $p(\mathsf{x}) = (1/M)\sum_{m=1}^M \delta(\mathsf{x} - \mathsf{x}_m)$. Note, however, that for high-dimensional feature spaces direct evaluation of the averages is implausible. It is therefore desirable to avoid explicit computations in $\{\boldsymbol{\phi}\}$.

3.1 Kernelized Representation

Since each row $\tilde{\mathsf{w}}_i^T \in \mathbb{R}^{1\times|\boldsymbol{\phi}|}$ of the weight matrix $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\boldsymbol{\phi}|}$ has the same dimensionality as the feature vectors $\boldsymbol{\phi}(\mathsf{x}_i)^T$, it is representable as

$$\tilde{\mathsf{w}}_i = \sum_{m=1}^{M} \alpha_{im}\boldsymbol{\phi}(\mathsf{x}_m) + \tilde{\mathsf{w}}_i^{\perp}, \tag{9}$$

where $\tilde{\mathsf{w}}_i^{\perp}$ is orthogonal to the span of $\phi(\mathsf{x}_1), \ldots, \phi(\mathsf{x}_M)$. Then

$$\mathsf{W} = \mathsf{A}\mathsf{F}^T + \mathsf{W}^{\perp}, \quad \mathsf{F} \overset{\text{def}}{=} [\boldsymbol{\phi}(\mathsf{x}_1), \ \ldots, \ \boldsymbol{\phi}(\mathsf{x}_M)] \in \mathbb{R}^{|\boldsymbol{\phi}| \times M}, \tag{10}$$

where $\mathsf{A} = \{\alpha_{ij}\} \in \mathbb{R}^{|\mathsf{y}| \times M}$ is the matrix of coefficients, and the transposed rows of $\mathsf{W}^{\perp}$ are given by $\tilde{\mathsf{w}}_i^{\perp}$. In kernel literature F is often referred to as the *design* matrix (e.g. [18]).

From (10), we obtain expressions for the averages

$$\begin{aligned}\mathsf{W}\langle\boldsymbol{\phi}(\mathsf{x})\mathsf{x}^T\rangle &= \mathsf{A}\left\langle \left[\boldsymbol{\phi}(\mathsf{x}_1)^T\boldsymbol{\phi}(\mathsf{x}), \ldots, \boldsymbol{\phi}(\mathsf{x}_M)^T\boldsymbol{\phi}(\mathsf{x})\right]^T \mathsf{x}^T \right\rangle \\ &= \frac{\mathsf{A}}{M}\left[\sum_m \mathsf{x}_m \boldsymbol{\phi}(\mathsf{x}_m)^T\boldsymbol{\phi}(\mathsf{x}_1), \ \ldots\right]^T = \frac{\mathsf{A}\mathsf{B}^T}{M}\end{aligned}$$

where we defined

$$\mathsf{B} \overset{\text{def}}{=} \sum_{m=1}^{M} \mathsf{x}_m \mathsf{k}(\mathsf{x}_m)^T \in \mathbb{R}^{|\mathsf{x}| \times M}, \ \ \mathsf{k}(\mathsf{x}_m) = \mathsf{F}^T\boldsymbol{\phi}(\mathsf{x}_m) \in \mathbb{R}^M. \tag{11}$$

Here $\mathsf{k}(\mathsf{x}_m)$ corresponds to the m^{th} column (or row) of the Gram matrix $\mathsf{K} = \{K_{ij}\} \overset{\text{def}}{=} \{\boldsymbol{\phi}(\mathsf{x}_i)^T\boldsymbol{\phi}(\mathsf{x}_j)\} \in \mathbb{R}^{M \times M}$. Finally, note that for a fixed K the computed expectation $\mathsf{W}\langle\boldsymbol{\phi}(\mathsf{x})\mathsf{x}^T\rangle$ is a function of A.

Analogously, we can express

$$\mathsf{W}\langle\boldsymbol{\phi}(\mathsf{x})\boldsymbol{\phi}(\mathsf{x})^T\rangle\mathsf{W}^T = \frac{1}{M}\mathsf{A}\sum_{m=1}^{M}\mathsf{k}(\mathsf{x}_m)\mathsf{k}(\mathsf{x}_m)^T\mathsf{A}^T. \tag{12}$$

Again, for the fixed Gram matrix the term is a quadratic function of coefficients A alone.

By substitution, we may re-express the bound (**??**) as

$$\tilde{I}(\mathsf{x},\mathsf{y}) = \frac{1}{M}\mathrm{tr}\left\{\boldsymbol{\Sigma}_x^{-1}\mathsf{U}\mathsf{A}\mathsf{B}^T\right\} - \frac{1}{2}\mathrm{tr}\left\{\mathsf{U}^T\boldsymbol{\Sigma}_x^{-1}\mathsf{U}\boldsymbol{\Sigma}_y\right\} - \frac{1}{2M}\mathrm{tr}\left\{\mathsf{U}^T\boldsymbol{\Sigma}_x^{-1}\mathsf{U}\mathsf{A}\mathsf{K}^2\mathsf{A}^T\right\} - \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_x^{-1}\mathsf{S}\right\}. \tag{13}$$

In the simplest case when $\boldsymbol{\phi}(\mathsf{x}) \equiv \mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$, we obtain $\mathsf{K}^2 \propto \mathsf{X}^T\mathsf{S}\mathsf{X} \in \mathbb{R}^{M \times M}$, $\mathsf{B} \propto \mathsf{S}\mathsf{X} \in \mathbb{R}^{|\mathsf{x}| \times M}$ where $\mathsf{X} \overset{\text{def}}{=} [\mathsf{x}_1, \ \ldots, \ \mathsf{x}_M] \in \mathbb{R}^{|\mathsf{x}| \times M}$ contains the training data. As expected, this transforms the bound (13) to the corresponding expression (7) for the linear Gaussian channel, thus resulting in PCA on the sample covariance S for both the encoder W^T and the decoder U as the optimal choice.

The objective (13) may be used to learn optimal decoder weights U and optimal coordinates A in the space spanned by the feature vectors $\{\boldsymbol{\phi}(\mathsf{x}_i)|i \in [1, M] \cap \mathbb{N}\}$. Moreover, note that if $\mathcal{K}_{\boldsymbol{\Theta}} : |\mathsf{x}| \times |\mathsf{x}| \to \mathbb{R}$ defines a symmetric positive-definite *kernel*, by Mercer's theorem the elements K_{ij} of the Gram matrix may be replaced by $\mathcal{K}(\mathsf{x}_i, \mathsf{x}_j; \boldsymbol{\Theta})$. In this case, the bound $\tilde{I}(\mathsf{x},\mathsf{y})$ may be used to learn the optimal parameters[1] of the kernel function. Thus, in our framework the procedure for learning kernels may be viewed as a special case of the variational IM algorithm [1].

3.2 Nature of optimal solutions

In the following we assume for simplicity that $\boldsymbol{\Sigma}_y = s^2\mathsf{I}$ and $\boldsymbol{\Sigma}_x = \sigma^2\mathsf{I}$. We also assume that $|\mathsf{y}| \leq |\mathsf{x}| \leq |\boldsymbol{\phi}|$ and $|\mathsf{x}| \leq M$, so that y is a compressed representation of $\boldsymbol{\phi}(\mathsf{x})$, and the number of training points is sufficient to ensure invertibility of the sample covariance.

[1] It is also possible to learn K directly by constraining it to satisfy properties of inner products; this may be useful, for example, when the source alphabet is exhausted by M training points.

Optimal Decoder

Optimization of $\tilde{I}(\mathsf{x},\mathsf{y})$ for the matrix of coefficients A leads to the fixed point condition

$$\partial\tilde{I}(\mathsf{x},\mathsf{y})/\partial\mathsf{A} = 0 \quad \Rightarrow \mathsf{U}^T\mathbf{\Sigma}_x^{-1}\mathsf{B} = \mathsf{U}^T\mathbf{\Sigma}_x^{-1}\mathsf{UAK}^2. \tag{14}$$

For a non-singular Gram matrix K we obtain

$$\tilde{I}(\mathsf{x},\mathsf{y}) \propto \mathrm{tr}\left\{\mathsf{UAK}^2\mathsf{A}^T\mathsf{U}^T\right\} - s^2\mathrm{tr}\left\{\mathsf{UU}^T\right\} = \mathrm{tr}\left\{\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{BK}^{-2}\mathsf{B}^T\right\} - s^2\mathrm{tr}\left\{\mathsf{UU}^T\right\}. \tag{15}$$

Let $\tilde{\mathsf{S}}_F = \sum_{m=1}^M \mathsf{x}_m\boldsymbol{\phi}(\mathsf{x}_m)^T/M \in \mathbb{R}^{|\mathsf{x}|\times|\boldsymbol{\phi}|}$. By noticing that

$$\mathsf{BK}^{-2}\mathsf{B}^T \propto \tilde{\mathsf{S}}_F\mathsf{FK}^{-2}\mathsf{F}^T\tilde{\mathsf{S}}_F^T \propto \mathsf{S}, \tag{16}$$

the bound (15) reduces to the objective of the linear Gaussian channel (6). Note that $\|\mathsf{W}\|_F \to \infty$ as $\|\mathsf{U}\|_F \to 0$.

Thus, under the specific assumption of isotropic Gaussian noise, optimal weights U of the linear Gaussian decoder correspond to principal components (and their rotations) of the sample covariance S. Similarly to the case of a linear channel, for $s^2 > 0$ it is necessary to impose norm constraints on U to ensure convergence of $\|\mathsf{W}\|_F$. Fundamentally, therefore, this simple decoder will restrict severely the power of the approach, as we will see below.

Optimal Encoder

From (14) it is clear that optimal solutions for the encoder are given by

$$\mathsf{WF} = \mathsf{AK} \propto \mathsf{U}^+\mathsf{X}, \tag{17}$$

where U^+ denotes the pseudo-inverse, and left singular values of U correspond to principal eigenvectors of S. In the case when $\boldsymbol{\phi}(\mathsf{x}) \equiv \mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$ and XX^T is non-singular, condition (17) results in $\mathsf{W} = \mathsf{U}^+ \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$, which is the PCA solution of the linear Gaussian channel. However, for general non-linear mappings, optimal encoder weights W^T do not necessarily give rise to the non-linear PCA solution.

Finally, note that if the channel noise is isotropic and there are no constraints preventing the weights from taking optimal solutions according to (15) and (17), then the bound is given by the summation of $|\mathsf{y}|$ principal eigenvalues of the sample covariance S. In this case the objective is invariant under the choice of non-linearity. Hence, we reach an important conclusion: *for isotropic channel noise, if the decoder is linear, nothing is gained by using a non-linear encoder* in the proposed variational settings. In other cases, for example when the channel noise is correlated or the encoder and decoder weights have structural constraints, optimal parameters of $\mathcal{K}_{\mathbf{\Theta}}$ may be obtained by maximizing (13).

These results are somewhat disappointing. In order to improve the power of the method, we need to consider both non-linear encoders and decoders. However, the naive method of using a non-linear decoder would typically result in intractable averages over y. In order to avoid this difficulty, we consider how to form decoding in the feature space.

4 Nonlinear Decoders and KPCA

The non-linear Gaussian channel discussed in Section 3 may be represented by the Markov chain $\mathsf{x} \to \mathsf{f} \to \mathsf{y}$, where $\mathsf{f} \in \mathbb{R}^{|\boldsymbol{\phi}|}$ and $p(\mathsf{f}|\mathsf{x}) \sim \boldsymbol{\delta}(\mathsf{f} - \boldsymbol{\phi}(\mathsf{x}))$, $p(\mathsf{y}|\mathsf{f}) \sim \mathcal{N}_y(\mathsf{Wf}, \mathbf{\Sigma}_y)$. Indeed, by marginalizing the feature variables f it is clear that the encoder is given[2] by $p(\mathsf{y}|\mathsf{x}) = \int_{\mathsf{f}} \boldsymbol{\delta}(\mathsf{f} - \boldsymbol{\phi}(\mathsf{x}))\mathcal{N}_y(\mathsf{Wf}, \mathbf{\Sigma}_y) = \mathcal{N}_y(\mathsf{W}\boldsymbol{\phi}(\mathsf{x}), \mathbf{\Sigma}_y)$.

[2] We assume Cartesian coordinates, i.e. $\boldsymbol{\delta}(\mathsf{x} - \mathsf{a}) = \prod_i \delta(x_i - a_i)$.

Proposition 1: *Let* $\mathsf{s} \to \mathsf{t} \to \mathsf{r}$ *define a Markov chain, such that* $p(\mathsf{t}|\mathsf{s}) = \delta(\mathsf{t} - \mathsf{f}(\mathsf{s}))$*, and* $p(\mathsf{r}|\mathsf{t})$ *is a continuous differentiable density function satisfying* $\forall \mathsf{r}.\forall \mathsf{t}.p(\mathsf{r}|\mathsf{t}) \neq 0$*. Then* $I(\mathsf{s},\mathsf{r}) = I(\mathsf{t},\mathsf{r})$.

From proposition 1, the mutual information $I(\mathsf{x},\mathsf{y})$ may be bounded as

$$I(\mathsf{x},\mathsf{y}) = I(\mathsf{f},\mathsf{y}) \geq \tilde{I}(\mathsf{f},\mathsf{y}), \text{ where } \tilde{I}(\mathsf{f},\mathsf{y}) \stackrel{\text{def}}{=} \langle \log q(\mathsf{f}|\mathsf{y}) \rangle_{p(\mathsf{x})p(\mathsf{f}|\mathsf{x})p(\mathsf{y}|\mathsf{f})} + H(\mathsf{f}). \tag{18}$$

We make the simple assumption that the feature decoder is Gaussian, $q(\mathsf{f}|\mathsf{y}) \sim \mathcal{N}_f(\mathsf{Uy}, \mathbf{\Sigma}_f)$, for which

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2}\text{tr}\left\{\mathsf{U}^T \mathbf{\Sigma}_f^{-1} \mathsf{U}\left(\mathbf{\Sigma}_y + \mathsf{W}\mathsf{S}_F\mathsf{W}^T\right)\right\} + \text{tr}\left\{\mathbf{\Sigma}_f^{-1}\mathsf{U}\mathsf{W}\mathsf{S}_F\right\} + H(\mathsf{f}) + c \tag{19}$$

where $\mathsf{S}_F \stackrel{\text{def}}{=} \langle \mathsf{f}\mathsf{f}^T \rangle_{p(\mathsf{f})}$. If $\langle \mathsf{f} \rangle = 0$ then clearly S_F corresponds to the sample covariance in the feature space (see [14] for a discussion of centering of the data in feature spaces). This covariance is readily computable from the training set as

$$\mathsf{S}_F \quad = \quad \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\phi}(\mathsf{x}_m)\boldsymbol{\phi}(\mathsf{x}_m)^T, \tag{20}$$

where we assumed that $\mathsf{x} = \boldsymbol{\phi}^{-1}(\mathsf{f})$. Note that if $\boldsymbol{\phi} : \mathsf{x} \to \mathsf{f}$ is deterministic and $p(\mathsf{x}) = \sum_{i=1}^{M} \boldsymbol{\delta}(\mathsf{x} - \mathsf{x}_i)/M$ then $\boldsymbol{\phi}(\mathsf{x}_i)$ corresponds to a re-labeling of the source pattern, leading to $H(\mathsf{f}) = H(\mathsf{x})$. As expected, linear mappings $\boldsymbol{\phi}(\mathsf{x}) \equiv \mathsf{x}$ result in $\mathsf{S}_F = \langle \mathsf{x}\mathsf{x}^T \rangle \equiv \mathsf{S}$, which reduces the objective (19) to (4).

It is easy to see that in general this case gives rise to non-linear source decoders $q(\mathsf{x}|\mathsf{y}) = \int_{\mathsf{f}} p(\mathsf{x}|\mathsf{f})\mathcal{N}_f(\mathsf{Uy}, \sigma_f^2\mathsf{I})$; however, they may be difficult to compute. This is an important limitation of the approach, since for general feature mappings $\boldsymbol{\phi}$ it may be difficult to reconstruct x from its encoded representation y.

4.1 Kernelized Representation

In what follows we assume that $\mathbf{\Sigma}_y = s^2\mathsf{I} \in \mathbb{R}^{|\mathsf{y}|}$, $\mathbf{\Sigma}_f = \sigma_f^2\mathsf{I} \in \mathbb{R}^{|\phi|}$, and $|\mathsf{y}| < M < |\boldsymbol{\phi}|$. By analogy with Section 3 we notice that rows of W and columns of U have dimension $|\boldsymbol{\phi}|$. Then they may be represented in the basis defined by the span of $\{\boldsymbol{\phi}(\mathsf{x}_m)|m = 1, \ldots, M\}$ and its orthogonal compliment as

$$\mathsf{W} = \mathsf{A}\mathsf{F}^T + \mathsf{W}^{\perp} \in \mathbb{R}^{|\mathsf{y}|\times|\boldsymbol{\phi}|}, \quad \mathsf{U} = \mathsf{F}\mathsf{C} + \mathsf{U}^{\perp} \in \mathbb{R}^{|\boldsymbol{\phi}|\times|\mathsf{y}|}. \tag{21}$$

Here $\mathsf{F} \in \mathbb{R}^{|\boldsymbol{\phi}|\times M}$ is the design matrix, $\mathsf{U}^{\perp}$, $\mathsf{W}^{\perp}$ are orthogonal to F, and $\mathsf{A} \in \mathbb{R}^{|\mathsf{y}|\times M}$, $\mathsf{C} \in \mathbb{R}^{M\times|\mathsf{y}|}$ are matrices of coefficients to be learned. Substitution into expression (19) results in

$$\tilde{I}(\mathsf{x},\mathsf{y}) \propto 2\text{tr}\left\{\mathsf{C}\mathsf{A}\mathsf{K}^2\right\} - s^2\text{tr}\left\{\mathsf{C}^T\mathsf{K}\mathsf{C}\right\} - \text{tr}\left\{\mathsf{A}\mathsf{K}^2\mathsf{A}^T\left[\mathsf{C}^T\mathsf{K}\mathsf{C} + (\mathsf{U}^{\perp})^T\mathsf{U}^{\perp}\right]\right\}, \tag{22}$$

where we ignored the terms independent of A, C and K. It is clear that since we are interested in maximizing the bound, we may assume $\mathsf{U}^{\perp} = \mathbf{0} \in \mathbb{R}^{|\boldsymbol{\phi}|\times|\mathsf{y}|}$.

The objective (22) may be readily used for learning coefficients A, C. In the considered case it may also be applied to learning parameters $\mathbf{\Theta}$ of the kernel function $\mathcal{K}(\mathsf{x}_i, \mathsf{x}_j; \mathbf{\Theta})$ which gives rise to the Gram matrix K.

4.2 Nature of Optimal Solutions

Optimization of the bound (22) for the coefficients A results in

$$\mathsf{A} = (\mathsf{C}^T\mathsf{K}\mathsf{C})^{-1}\mathsf{C}^T, \tag{23}$$

which transforms the objective to

$$\tilde{I}(\mathsf{x},\mathsf{y}) = \mathrm{tr}\left\{\mathsf{K}^2\mathsf{C}(\mathsf{C}^T\mathsf{K}\mathsf{C})^{-1}\mathsf{C}^T\right\} - s^2\mathrm{tr}\left\{\mathsf{C}^T\mathsf{K}\mathsf{C}\right\} = \mathrm{tr}\left\{\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{S}_F\right\} - s^2\mathrm{tr}\left\{\mathsf{U}\mathsf{U}^T\right\} \tag{24}$$

By analogy with Section 2, maximization of (24) gives rise to the non-linear PCA solution (and rotations) for the left singular vectors of U and W^T. (As before, we have ignored $\mathsf{W}^\perp$ and $\mathsf{U}^\perp$ in the definitions (21) since they have no influence on the bound).

Just as in the linear case, in order to prevent divergence of $\|\mathsf{W}\|_F$ for $s^2 \neq 0$, it is useful to constrain the singular values of U. In the special case when $\mathsf{U}^T\mathsf{U} = \mathsf{I}$, expressions (21) and (24) lead to

$$\mathsf{K}^2\mathsf{C}\mathsf{R} = \mathsf{K}\mathsf{C}\mathsf{R}\mathbf{\Lambda}_{S_F}. \tag{25}$$

Here $\mathbf{\Lambda}_{S_F} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is a diagonal matrix of eigenvalues of S_F and $\mathsf{R} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is a rotation matrix. From (23) and (25) it is clear that optimal C and A^T correspond to rotations of principal eigenvectors of the Gram matrix K, which is the kernel PCA solution. Hence, *kernel PCA can be viewed as a lower bound on the mutual information between* x *and* y *in nonlinear Gaussian channels.* Indeed, the bound enables us, in a principled way, to choose between different parameters of the kernel function, or to choose between competing kernels.

5 Optimal Kernel Functions

Optimal parameters $\mathbf{\Theta}$ of the kernel function $\mathcal{K}_{\mathbf{\Theta}} : |\mathsf{x}| \times |\mathsf{x}| \to \mathbb{R}$ may be obtained by maximizing the general objectives (13) or (19) or their kernelized versions. The optimization procedure may be viewed as a special case of the IM algorithm:

1. For the fixed Gram matrix K, optimize the bound w.r.t U, W (or the dual parameters C and A).
2. For the fixed U, W (or C and A), optimize the bound w.r.t. the parameters $\mathbf{\Theta}$ of $\mathcal{K}_{\mathbf{\Theta}}$.

For non-isotropic channels (or constrained encoder-decoder pairs) this may in general result in non-trivial settings of the parameters $\mathbf{\Theta}$. In what follows we describe a few special cases of optimal kernel functions for the simplest KPCA channel of Section 4.2. Our motivation here is gaining an intuition into choosing kernel parameters which maximize the bound on the mutual information (see [15], [16] for a detailed discussion of concentration properties of Gram matrices).

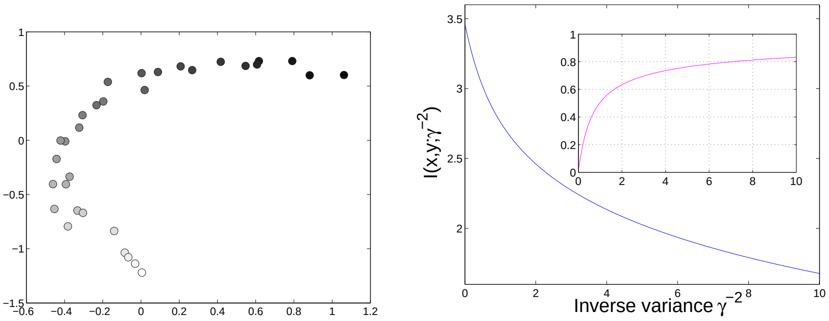
5.1 Kernel PCA Channels

As before, we assume that $\mathsf{K} \in \mathbb{R}^{M\times M}$ is non-singular. From (24) and (25) it is clear that an alternative formulation of the information maximization problem for the KPCA case is given by

$$\max_{\mathbf{\Theta}} \max_{\mathsf{C},\mathsf{M}} \left[\mathrm{tr}\left\{\mathsf{C}^T\mathsf{K}\mathsf{C}\right\} - \mathrm{tr}\left\{\mathsf{M}(\mathsf{C}^T\mathsf{C} - \mathsf{I})\right\}\right] \equiv \max_{\mathbf{\Theta}} \sum_{i=1}^{|\mathsf{y}|} \lambda_i(\mathcal{K}_{\mathbf{\Theta}}) \leq \mathrm{tr}\left\{\mathsf{K}\right\}. \tag{26}$$

Here $\mathsf{M} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is a matrix of Lagrange multipliers imposing orthonormality constraints on $\mathsf{C} \in \mathbb{R}^{M\times|\mathsf{y}|}$, and $\lambda_i(\mathcal{K}_{\mathbf{\Theta}})$ is the i^{th} principal component of the Gram matrix $\mathsf{K} \in \mathbb{R}^{M\times M}$ corresponding to $\mathcal{K}_{\mathbf{\Theta}}$.

In general, in order to avoid divergence it is necessary to impose norm constraints on $\|\mathsf{K}\|_F$. In this case it is intuitive that the worst kernel has a flat spectrum (which is the case for $\mathsf{K} = c\mathsf{I}$), while the optimal kernel function results in the Gram matrix K with an eigenspectrum concentrated at $|\mathsf{y}|$ principal components.



(a) (b)

Figure 1: *(a):* training data, $M = 50$, $|\mathsf{x}| = 2$. *(b):* change of $\tilde{I}(\mathsf{x}, \mathsf{y}; \gamma^2)$ with an increase of the *inverse* variance γ^{-2} for Squared Exponential kernels, $|\mathsf{y}| = 1$. *Insert:* Residuals $\mathrm{res}(\gamma^{-2}) \overset{\mathrm{def}}{=} (M - \lambda_1(\mathcal{K}_\gamma))/M$ as a function of γ^{-2}.

Squared Exponential (RBF) Kernels

Let $\mathcal{K}_\gamma(\mathsf{x}_i, \mathsf{x}_j; \gamma^2) \overset{\mathrm{def}}{=} \exp\{-\|\mathsf{x}_i - \mathsf{x}_j\|^2/(2\gamma^2)\}$ be a kernel function with the variance γ^2. It is clear that $K_{ij} \leq K_{ii} = 1$ and $\mathrm{tr}\,\{\mathsf{K}\} = M$.

Proposition 2: *For Squared Exponential kernels $\mathcal{K}_\gamma(\mathsf{x}, \tilde{\mathsf{x}}; \gamma^2)$ the optimal solution of (26) is attained in the limit of diverging variance γ^2.*

Mixture Kernels

If $\mathcal{K}_1$ and $\mathcal{K}_2$ are positive semi-definite kernel functions then so is $\mathcal{K}_\alpha \overset{\mathrm{def}}{=} \alpha\mathcal{K}_1 + (1 - \alpha)\mathcal{K}_2$, where $\alpha \in [0, 1]$. Here we show that $\mathcal{K}_\alpha$ optimized for the mixing coefficient α converges to the single best kernel component. This results from the following propositions:

Proposition 3: *Let $\mathsf{L} = \mathrm{diag}(\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^{M \times M}$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$, and $\mathsf{R} \in \mathbb{R}^{M \times |\mathsf{y}|}$ such that $\mathsf{R}^T\mathsf{R} = \mathsf{I}_{|\mathsf{y}|}$. Then $\mathrm{tr}\left\{\mathsf{R}^T\mathsf{L}\mathsf{R}\right\} \leq \sum_{i=1}^{|\mathsf{y}|} \lambda_i$.*

Proposition 4: *Let $\mathcal{K}_\alpha \overset{\mathrm{def}}{=} \alpha\mathcal{K}_1 + (1 - \alpha)\mathcal{K}_2$, where $\alpha \in [0, 1]$. Then the optimal solution of (26) is obtained for $\alpha \in \{0, 1\}$.*

The obtained result is quite intuitive. Just as a mixture distribution is generally vaguer than a distribution defined by a single mixture component, eigenspectra of mixtures of positive semi-definite matrices are generally flat. This makes them sub-optimal in the KPCA channels under the objective (26). Finally, note that the results of proposition 4 may be easily generalized to any convex mixture of fixed kernel functions.

6 Experiments

Visualizing a Low-Dimensional Structure: Let $\mathcal{K}_\gamma(\mathsf{x}_i, \mathsf{x}_j; \gamma^2)$ be a Squared Exponential kernel with a fixed variance parameter γ^2. Here we show the influence of γ^2 on the bound (22) for a simple KPCA channel (see Section 4.2). We also discuss performance of kernels with various parameter settings for visualizing an intrinsically low-dimensional structure of the data.

Figure 1 *(a)* shows $M = 50$ training points $\mathsf{x}_i \in \mathbb{R}^2$. The points were sampled uniformly from a semicircular unit arc, and perturbed by a small noise $\epsilon \sim \mathcal{N}_\epsilon(\mathbf{0}, 0.05\mathsf{I}_2)$. From the construction it is clear that the (approximate) intrinsic dimension of each data point x_i is defined by an angle $\psi(\mathsf{x}_i) \in \mathbb{R}$. One may hope that maximization of $I(\mathsf{x}, y)$ may result in the codes which capture useful 1d information about the angles. Since the points were sampled uniformly and the noise was small, the phase change $\Delta\psi(\mathsf{x}_{i+1}, \mathsf{x}_i) \overset{\text{def}}{=} \psi(\mathsf{x}_{i+1}) - \psi(\mathsf{x}_i)$ between the neighboring points should be approximately constant. If $y(\mathsf{x}_i)$ indeed corresponds to a (scaled) phase $\psi(\mathsf{x}_i)$, we can expect $y(\mathsf{x}_i)$ to be roughly linear as a function of i.

Figure 1 *(b)* shows the change in $\log \tilde{I}(\mathsf{x}, \mathsf{y}; \gamma^2)$ with a decrease in the variance. We see that the maximum of the bound is attained in the limit of $\gamma^2 \to \infty$ (*cf* proposition 2). The inserted graph shows the relative weight of the $M-1$ minor eigenvalues $\sum_{i=2}^{M} \lambda_i(\mathcal{K}_\gamma)/M$, which decreases as $\gamma^2 \to \infty$. We therefore expect that in the specified channel the information transmission improves with the growth of γ^2.

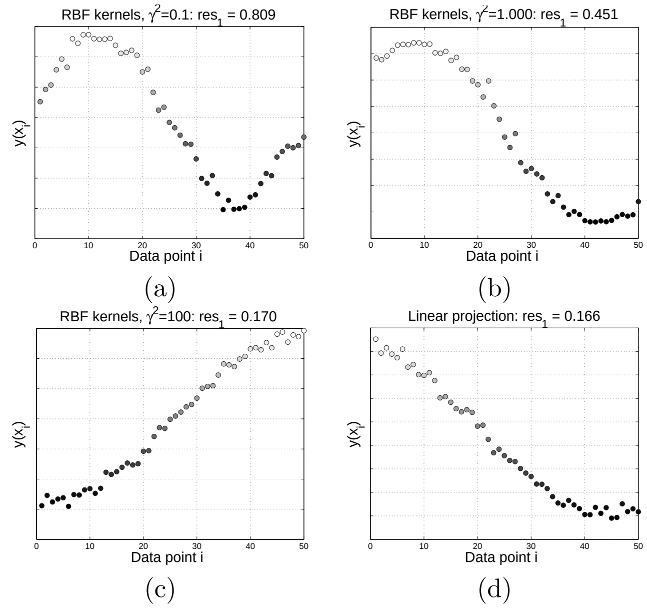
To confirm that the choice of γ^2 influences the visualization performance, we sampled $M_1 = 75$ testing points at uniform from the same process. Figure 2 plots projections y_i of the testing points x_i as a function of i (for illustration purposes, higher values of y_i are plotted in lighter colors). For a better visualization, the data was centered in the feature space (see [14]). Parameters γ^2 of the kernel functions $\mathcal{K}_\gamma$ were fixed at $\gamma^2 = 0.1, 1$, and 100, which resulted in the residuals decreasing as $\text{res}(\gamma^2) \approx 0.809, 0.446, 0.170$. From figure 2 *(a), (b), (c)* we see that as γ^2 increases, $y(\mathsf{x}_i)$ becomes approximately linear in i. For a high γ^2 this indicates an approximately linear change of projection coordinates, which is characteristic of the intrinsic parameterization of the data. Finally, figure 2 *(d)* shows a linear projection of the testing data onto the principal eigenvector of the sample covariance S. One can see that even though a part of the plot is approximately linear, $\psi(\mathsf{x}_i)$ has an extraneous region of the constant phase towards the darker end of the line (which is easily explained by the form of the training data shown on figure 1 *(a)*).

Digit Clustering: Figure 3 shows projections of the real-world digits onto the 3 principal components of the feature covariance S_F for the squared exponential kernel function $\mathcal{K}_\gamma$, $\gamma^2 = 100$. The training data consisted of $M = 90$ instances of digits 4, 5, and 7 (30 of each kind) with $|\mathsf{x}| = 196$. Prior to clustering, the patterns were centered and normalized. It is intuitive that if y contains information about distances to cluster centers, it may be useful for reconstructing x, so clustering is easily explainable. However, this is mainly an effect of constraints on the channel parameters (as will be discussed elsewhere).

7 Discussion

We have shown that optimal Gaussian decoders have a strong relation to popular dimensionality reduction techniques. One of the main contributions of this paper is a principled objective function which can be used to learn and compare kernel parameters. One can also envisage a similar approach based on a kernelized version of autoencoders. In particular, the bound here for the Gaussian decoders is strongly related to using a minimal least squares linear reconstruction error. However, our method is arguably more general, since it naturally generalizes to different channels, and is well founded from an information theoretic viewpoint. The result that using a linear decoder does not improve the situation much is analogous to the well-known result in autoencoders that one needs more than one hidden unit to improve on PCA [3].

Recently, Lawrence has introduced a nice way to use Gaussian processes for dimension reduction [9]. Our work is related to this, in the sense that both methods produce PCA as a limiting special case. However, the work of Lawrence is not directly related to kernel PCA, although it does indeed enable the evaluation of different kernel functions. From a practical viewpoint, our method may be computationally rather more convenient since, unlike [9], it does not require non-linear optimization to perform dimensionality reduction.



(a)

(b)

(c)

(d)

Figure 2: Projection $y(\mathsf{x}_i)$ as a function of i for $M_1 = 75$, $|\mathsf{y}| = 1$. *(a), (b), (c):* RBF kernels, $\gamma^2 = 0.1, 1, 100$. *(d):* A linear projection onto the 1^{st} principal eigenvector of S.

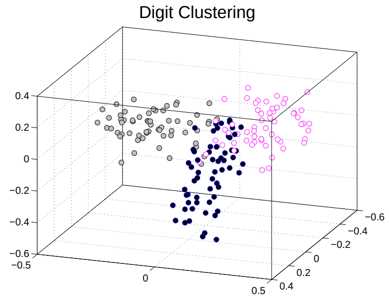


Figure 3: Clustering of 180 instances of 3 digits with RBF kernels, $|\mathsf{x}| = 196$, $|\mathsf{y}| = 3$, $\gamma^2 = 100$.

Appendix

Proposition 1 : Proof: From basic properties of the mutual information (see e.g. [5]) it is easy to see that

$$\begin{aligned} I(\mathsf{s},\mathsf{t};\mathsf{r}) &= H(\mathsf{r}) - H(\mathsf{r}|\mathsf{t},\mathsf{s}) = I(\mathsf{t},\mathsf{r}), && (27)\\ I(\mathsf{s},\mathsf{t};\mathsf{r}) &= H(\mathsf{s}) + H(\mathsf{t}|\mathsf{s}) - H(\mathsf{s}|\mathsf{r}) - H(\mathsf{t}|\mathsf{s},\mathsf{r}) \\ &= I(\mathsf{s},\mathsf{r}) + H(\mathsf{t}|\mathsf{s}) - H(\mathsf{t}|\mathsf{s},\mathsf{r}). && (28)\end{aligned}$$

Utilizing the chain structure and the deterministic mapping $p(\mathsf{t}|\mathsf{s})$, we obtain

$$p(\mathsf{t}|\mathsf{s},\mathsf{r}) = \frac{\boldsymbol{\delta}(\mathsf{t}-\mathsf{f}(\mathsf{s}))p(\mathsf{r}|\mathsf{t})}{\int_{\mathsf{t}}\boldsymbol{\delta}(\mathsf{t}-\mathsf{f}(\mathsf{s}))p(\mathsf{r}|\mathsf{t})} = \frac{\boldsymbol{\delta}(\mathsf{t}-\mathsf{f}(\mathsf{s}))p(\mathsf{r}|\mathsf{f}(\mathsf{s}))}{p(\mathsf{r}|\mathsf{f}(\mathsf{s}))}, \tag{29}$$

i.e. $p(\mathsf{t}|\mathsf{s},\mathsf{r}) = p(\mathsf{t}|\mathsf{s})$. Here we used $f(\mathsf{x})\delta(\mathsf{x}-\mathsf{a}) = \lim_{\mathsf{e}\to 0}\left[f(\mathsf{a}-\mathsf{e}) + f(\mathsf{a}+\mathsf{e})\right]\delta(\mathsf{x}-\mathsf{a})/2$ (see e.g. [8]). Then $H(\mathsf{t}|\mathsf{s},\mathsf{r}) = H(\mathsf{t}|\mathsf{s})$, and from (27), (28) we obtain $I(\mathsf{s},\mathsf{r}) = I(\mathsf{t},\mathsf{r})$. ■

Proposition 2 : Proof: From the definition of $\mathcal{K}_\gamma(\mathsf{x}, \tilde{\mathsf{x}}; \gamma^2)$, we get $\forall i. \forall j. \lim_{\gamma^2 \to \infty} K_{ij} = 1$, i.e. rank(K) → 1 as $\gamma^2 \to \infty$. Then $\lim_{\gamma^2 \to \infty} \sum_{i=1}^{|\mathsf{y}|} \lambda_i(\mathcal{K}_\gamma) = \lambda_1(\mathcal{K}_\gamma) = M = \mathrm{tr}\{\mathsf{K}\}$, which is a global optimum of the objective in (26) independently of the size of $|\mathsf{y}|$. ■

Proposition 3 : Proof: The proof follows straight away from the solution of the constrained optimization problem $\tilde{\mathsf{R}} = \mathrm{argmax}_{\mathsf{R}} \hat{I}_{\mathsf{R}}$, where

$$\hat{I}_{\mathsf{R}} = \mathrm{tr}\left\{\mathsf{R}^T \mathsf{L} \mathsf{R}\right\} - \mathrm{tr}\left\{\mathsf{M}(\mathsf{R}^T \mathsf{R} - \mathsf{I}_{|\mathsf{y}|})\right\}. \tag{30}$$

The optimization results in $\tilde{\mathsf{R}} \in \mathbb{R}^{M \times |\mathsf{y}|}$ corresponding to $|\mathsf{y}|$ principal components of $\mathsf{L} \in \mathbb{R}^{M \times M}$. Since L is a diagonal matrix with a sorted eigenspectrum, we get $\tilde{\mathsf{R}}^T = [\mathsf{I}_{|\mathsf{y}|}\ \mathbf{0}]$, where $\mathbf{0} \in \mathbb{R}^{|\mathsf{y}| \times (M - |\mathsf{y}|)}$ is a matrix of zeros. Then $\hat{I}_{\mathsf{R}} \leq \hat{I}_{\tilde{\mathsf{R}}} = \sum_{i=1}^{|\mathsf{y}|} \lambda_i$. ■

Proposition 4 : Proof:

Let $\mathsf{K}_\alpha = \mathsf{U} \mathbf{\Lambda} \mathsf{U}^T$, $\mathsf{K}_1 = \mathsf{U}_1 \mathbf{\Lambda}_1 \mathsf{U}_1^T$, and $\mathsf{K}_2 = \mathsf{U}_2 \mathbf{\Lambda}_2 \mathsf{U}_2^T$ be eigenvalue decompositions of the Gram matrices corresponding to the kernel functions $\mathcal{K}_\alpha$, $\mathcal{K}_1$, and $\mathcal{K}_2$. We are interested in maximizing $\sum_{i=1}^{|\mathsf{y}|} \lambda_i(\mathcal{K}_\alpha)$, which may be written as

$$\begin{aligned} \sum_{i=1}^{|\mathsf{y}|} \lambda_i(\mathcal{K}_\alpha) &= \mathrm{tr}\left\{\mathsf{V}^T \mathsf{U} \mathbf{\Lambda} \mathsf{U}^T \mathsf{V}\right\} \\ &= \alpha \mathrm{tr}\left\{\mathsf{R}_1^T \mathbf{\Lambda}_1 \mathsf{R}_1\right\} + (1-\alpha) \mathrm{tr}\left\{\mathsf{R}_2^T \mathbf{\Lambda}_2 \mathsf{R}_2\right\}. \end{aligned} \tag{31}$$

Here columns of $\mathsf{V} \in \mathbb{R}^{M \times |\mathsf{y}|}$ correspond to the principal eigenvectors of K_α, $\mathsf{R}_1^T \overset{\mathrm{def}}{=} \mathsf{V}^T \mathsf{U}_1 \in \mathbb{R}^{|\mathsf{y}| \times M}$, and $\mathsf{R}_2^T \overset{\mathrm{def}}{=} \mathsf{V}^T \mathsf{U}_2 \in \mathbb{R}^{|\mathsf{y}| \times M}$. From (31) and proposition 3 we get

$$\sum_{i=1}^{|\mathsf{y}|} \lambda_i(\mathcal{K}_\alpha) \leq \alpha \sum_{i=1}^{|\mathsf{y}|} \lambda_i^{(1)} + (1-\alpha) \sum_{i=1}^{|\mathsf{y}|} \lambda_i^{(2)} \leq \max_j \sum_{i=1}^{|\mathsf{y}|} \lambda_i^{(j)}, \tag{32}$$

where $\lambda_i^{(j)}$ corresponds to the i^{th} diagonal element of $\mathbf{\Lambda}_j$, $j \in \{1, 2\}$. Apart from the invariant cases, the maximum is achieved for $\alpha = 2 - \mathrm{argmax}_{j \in \{1,2\}} \sum_{i=1}^{|\mathsf{y}|} \lambda_i^{(j)} \in \{0, 1\}$. ■

References

[1] D. Barber and F. V. Agakov. The IM Algorithm: A Variational Approach to Information Maximization. In *NIPS*. MIT Press, 2003.

[2] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[3] C.M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[4] N. Brunel and J.-P. Nadal. Mutual Information, Fisher Information and Population Coding. *Neural Computation*, 10:1731–1757, 1998.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[6] N. Brunel J.-P. Nadal and N. Parga. Nonlinear feedforward networks with stochastic outputs: infomax implies redundancy reduction. *Network: Computation in Neural Systems*, 9(2):207–217, 1998.

[7] T. S. Jaakkola and M. I. Jordan. Improving the Mean Field Approximation via the Use of Mixture Distributions. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.

[8] G. A. Korn and T. M. Korn. *Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review*. McGraw-Hill, New York, 1968.

[9] N. D. Lawrence. Gaussian Process Latent Variable Models for Visualization of High Dimensional Data. In *NIPS*, 2003.

[10] R. Linsker. An Application of the Principle of Maximum Information Preservation to Linear Systems. In David Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1989.

[11] R. Linsker. How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation*, 1, 1989.

[12] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[13] D. Saad and M. Opper. *Advanced Mean Field Methods Theory and Practice*. MIT Press, 2001.

[14] B. Schoelkopf, A. Smola, and K.R. Mueller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10, 1998.

[15] J. Shawe-Taylor, N. Cristianini, and J. Kandola. On the Concentration of Spectral Properties. In *NIPS*. MIT Press, 2002.

[16] J. Shawe-Taylor and C. K. I. Williams. The Stability of Kernel Principal Components Analysis and its Relation to the Process Eigenspectrum. In *NIPS*. MIT Press, 2003.

[17] K. Torkkola and W. M. Campbell. Mutual Information in Learning Feature Transformations. *Proc. 17th International Conf. on Machine Learning*, 2000.

[18] C. K. I. Williams. Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.