Matching Kernels through Kullback-Leibler Divergence Minimisation

Neil D. Lawrence and Guido Sanguinetti
neil@dcs.shef.ac.uk, guido@dcs.shef.ac.uk

16th May 2005

Department of Computer Science, University of Sheffield, 211 Portobello Street, Sheffield, S1 4DP, U.K.

Abstract

In this paper we study the general constrained minimisation of Kullback-Leibler (KL) divergences between two zero mean Gaussian distributions. We reduce the problem to an equivalent minimisation involving the eigenvectors of the two kernel matrices, and provide explicit solutions in some cases.

We then focus, as an example, on the important case of constraining the approximating matrix to be block diagonal. We prove a stability result on the approximating matrix, and speculate on how these results may be used to give further theoretical foundation to widely used techniques such as spectral clustering.

1 Introduction

Kullback-Leibler (KL) divergence occupies a central role in statistical theory. It plays a fundamental role in variational inference techniques, and it is an objective function for a number of important a widely used models, such as probabilistic principal component analysis (PCA), kernel PCA and Gaussian process latent variable models [3, 2].

In this paper we study in detail the constrained minimisation of KL divergence. We assume that we have data, $\mathbf{Y}$, from which a positive (semi-)definite kernel matrix $\mathbf{K}_y$ is derived[1]. We seek to explain the data using latent variables, $\mathbf{X}$, which gives rise to a positive (semi-)definite matrix kernel $\mathbf{K}_x$. Depending on how $\mathbf{K}_x$ is constructed from $\mathbf{X}$ the principal sub-space of $\mathbf{K}_x$ may or may not be constrained. For example if $\mathbf{K}_x = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ then the sub-space of $\mathbf{K}_\mathbf{x}$ has a maximum rank of q if $\mathbf{X} \in \Re^{N\times q}$. On the other hand if $\mathbf{K}_\mathbf{x}$ is an RBF kernel based on $\mathbf{X}$ then $\mathbf{K}_x$ will have full rank. Because these matrices are positive definite, they can be viewed as correlation matrices, or covariance matrices.

We wish to 'match' in some sense the two correlation matrices, $\mathbf{K}_x$ and $\mathbf{K}_y$, thereby implying a match between $\mathbf{X}$ and $\mathbf{Y}$. It seems natural to minimise the Kullback-Leibler (KL) divergence [1] between the two distributions associated with these correlation matrices. This is a KL divergence between two multivariate Gaussians with zero mean.

In Section 2 we introduce the KL divergence between two Gaussian distributions as a function of the kernel parameters of the approximating distribution. We rewrite the KL divergence in terms of the singular value decomposition of the approximating kernel, and prove the equivalence of the KL minimisation problem with an easier optimisation problem involving only the scalar products between eigenvectors of the two kernel matrices. This provides deeper insight about the geometric properties of the optimal solution, and allows in certain cases to identify the optimum explicitly.

[1] This could be in the form of a kernel, a covariance matrix an affinity matrix *etc.*.

In Section 3 we apply our theoretical result to a simple constraint, when the approximating kernel is constrained to be block diagonal with each block of rank one (plus a spherical term to ensure positive definiteness). We prove that the solution is robust under perturbations.

Finallly, we discuss possible application of our results: in particular, the example of Section 3 has striking similarities with the kind of techniques used in spectral clustering. However, rigorous proof of a connection is not straigtforward and is beyond the scope of this paper.

2 Kullback-Leibler Divergence

In this section we will study the optimum of the Kullback-Leibler divergence with respect to the matrix $\mathbf{K}_x$. We shall proceed by considering eigendecompositions of $\mathbf{K}_x$ and $\mathbf{K}_y$. In our investigation we will generalise the proof given for the optimum in [3] which is in turn based on the proof given by [5] for the maximum likelihood solution of probabilistic PCA.

The Kullback-Leibler divergence between two Gaussian distributions,

$$\mathrm{KL}\left(N_y|N_x\right) = -\int N_y\left(\mathbf{z}|0,\mathbf{K}_y\right)\ln\frac{N_x\left(\mathbf{z}|0,\mathbf{K}_x\right)}{N_y\left(\mathbf{z}|0,\mathbf{K}_y\right)}d\mathbf{z},$$

for the special case of zero-mean Gaussians is given by,

$$\mathrm{KL}\left(N_y|N_x\right) = -\frac{1}{2}\log|\mathbf{K}_y| + \frac{1}{2}\log|\mathbf{K}_x| + \frac{1}{2}\mathrm{tr}\left(\mathbf{K}_y\mathbf{K}_x^{-1}\right) - \frac{N}{2}, \tag{1}$$

where we have taken the expectation under $N_y\left(\mathbf{z}|0,\mathbf{K}_y\right)$ and both correlation matrices are assumed to have dimension $N \times N$. The Kullback-Leibler divergence is asymmetric; we have chosen to consider the expectations under the distribution governed by the data correlation.

Covariance matrices are typically considered to be positive definite: to avoid problems arising from positive semi-definite kernels (which may be viewed as covariance matrices) we represent

$$\mathbf{K}_x = \mathbf{K}'_x + \beta^{-1}\mathbf{I}$$

and

$$\mathbf{K}_y = \mathbf{K}'_y + \tau^{-1}\mathbf{I}$$

where $\mathbf{K}'_x$ and $\mathbf{K}'_y$ can be positive semi-definite. In this way our results can also be viewed in the limit as τ or $\beta \to \infty$ if we are interested in positive semi-definite kernels.

2.1 Analysis

Our first step is to consider the eigendecomposition of $\mathbf{K}'_x$,

$$\mathbf{K}'_x = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathrm{T}}, \tag{2}$$

where $\mathbf{V}$ is a matrix of eigenvectors and $\mathbf{D}$ a diagonal matrix of eigenvalues. In the case of dimensional reduction, $\mathbf{D}$ may be of reduced rank, hence singular; however, the spherical term $\beta^{-1}\mathbf{I}$ will ensure positive definiteness. The inverse of $\mathbf{K}_x$ now takes the form

$$\mathbf{K}_x^{-1} = \mathbf{V}\left[\mathbf{D} + \beta^{-1}\mathbf{I}\right]^{-1}\mathbf{V}^{\mathbf{T}}$$

and the determinant is given by

$$|\mathbf{K}_x| = \prod_{i=1}^{N}\left(d_i + \beta^{-1}\right)$$

where d_i is the i-th element from the diagonal of $\mathbf{D}$. Through application of the Woodbury matrix inversion formula to $\mathbf{K}_x^{-1}$ the KL divergence may be rewritten

$$\begin{aligned}\mathrm{KL}\left(N_y|N_x\right) &= \frac{1}{2}\sum_{i=1}^{N}\log\left(d_i + \beta^{-1}\right) + \\ &\quad + \frac{1}{2}\mathrm{tr}\left(\mathbf{V}\left[\mathbf{D} + \beta^{-1}\mathbf{I}\right]^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{K}_y\right) + c\end{aligned}$$

where $c = -\frac{1}{2}\log|\mathbf{K}_y| - \frac{N}{2}$. We now make use of the fact that trace can be re-written as

$$\mathrm{tr}\left(\mathbf{V}\left[\mathbf{D}+\beta^{-1}\mathbf{I}\right]^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{K}_y\right) = \sum_{i=1}^{N}\hat{d}_i\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i$$

where $\hat{d}_i = \frac{1}{d_i+\beta^{-1}}$ is the i-th element from the diagonal of the matrix $\left[\beta^{-1}\mathbf{D}^{-1}+\mathbf{I}\right]^{-1}$ and $\mathbf{v}_i$ is the i-th column of $\mathbf{V}$. Substituting back into the KL divergence gives

$$\mathrm{KL}\left(N_y|N_x\right) = \frac{1}{2}\sum_{i=1}^{N}\log\left(d_i+\beta^{-1}\right) + \frac{1}{2}\sum_{i=1}^{N}\hat{d}_i\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i + c. \tag{3}$$

2.1.1 Solving for the Eigenvalues of $\mathbf{K}_x$

If we assume that the eigenvalues of $\mathbf{K}_x$ are only constrained by being greater or equal to zero then the solution with respect to the eigenvalues of $\mathbf{K}_x$ can be found by differentiating with respect to the eigenvalues of $\mathbf{K}_x$ giving

$$\frac{\partial \mathrm{KL}}{\partial d_i} = \frac{1}{2\left(d_i+\beta^{-1}\right)} - \frac{1}{2\left(\beta^{-1}+d_i\right)^2}\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i, \tag{4}$$

which implies that at a fixed point

$$d_i = \mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i - \beta^{-1} \text{ for } \beta^{-1} \leq \mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i. \tag{5}$$

We assume that only the first q entries in $\mathbf{D}$ are non-zero. The previous formula implies that for $i = 1 \ldots q$

$$\hat{d}_i = \frac{1}{\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i}$$

while for $i = q+1 \ldots N$ one has

$$\hat{d}_i = \beta.$$

Substituting this solution back into the KL divergence gives

$$\begin{aligned}\mathrm{KL}\left(N_y|N_x\right) &= \frac{1}{2}\sum_{i=1}^{q}\log\left(\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i\right) - \frac{N-q}{2}\log\beta \\ &\quad + \frac{\beta}{2}\sum_{i=q+1}^{N}\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i + + \frac{q}{2} + c\end{aligned} \tag{6}$$

where we have ordered the eigenvectors, $\{\mathbf{v}_i\}_{i=1}^{N}$, according to the size of $\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i$ and for the first q we have $\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i > \beta^{-1}$. This ordering holds without loss of generality.

2.1.2 Solving for β

We can now seek a solution for the value of β, again assuming that it is unconstrained in the positive half space. Differentiating (6) with respect to β gives

$$\frac{\partial \mathrm{KL}}{\partial \beta} = -\frac{N-q}{2\beta} + \frac{1}{2}\mathrm{tr}\left(\mathbf{K}_y\right) - \frac{1}{2}\sum_{i=1}^{q}\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i,$$

which implies that a fixed point for β is

$$\beta = \frac{N-q}{\mathrm{tr}\left(\mathbf{K}_y\right) - \sum_{i=1}^{q}\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i} \tag{7}$$

substituting into (6) we have

$$\begin{aligned}\mathrm{KL}\left(N_y|N_x\right) &= \frac{1}{2}\sum_{i=1}^{q}\log\left(\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i\right) - \frac{1}{2}\log|\mathbf{K}_y| \\ &\quad + \frac{N-q}{2}\log\left(\frac{\mathrm{tr}\left(\mathbf{K}_y\right) - \sum_{i=1}^{q}\mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i}{N-q}\right). \end{aligned} \tag{8}$$

2.1.3 Sub-space Matching

If we also considered the eigenvectors of $\mathbf{K}_x$ to be constrained only in number, not in direction, we could then proceed as in [5] and [3] to minimise the KL divergence. However, we wish to consider the more general case where the directions of the eigenvectors are constrained. To this end we consider the eigendecomposition of $\mathbf{K}'_y$. Recall that we defined

$$\mathbf{K}_y = \mathbf{K}'_y + \tau^{-1}\mathbf{I}$$

we write the eigendecomposition of $\mathbf{K}_y$ as

$$\mathbf{K}'_y = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}}$$

this enables us to write

$$\begin{aligned}\hat{a}_i &\triangleq \mathbf{v}_i^{\mathrm{T}}\mathbf{K}_y\mathbf{v}_i \\ &= \mathbf{v}_i^{\mathrm{T}}\mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}}\mathbf{v}_i + \tau^{-1} \\ &= \sum_{j=1}^{N}\lambda_j\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{v}_i\right)^2 + \tau^{-1} \\ &= \sum_{j=1}^{N}\lambda_j m_{ij} + \tau^{-1}\end{aligned} \tag{9}$$

where $m_{ij} = \left(\mathbf{u}_j^{\mathrm{T}}\mathbf{v}_i\right)^2$. Note that the matrix $\mathbf{M}$, which has elements m_{ij}, is 'doubly stochastic' ($\sum_{j=1}^{N} m_{ij} = 1$ and $\sum_{i=1}^{N} m_{ij} = 1$). The scalars m_{ij} encode the degree of alignment between the eigenvectors of $\mathbf{K_x}$ and those of $\mathbf{K}_y$.

2.1.4 KL Divergence as function of λ_j and m_{ij}.

We substitute this representation into (8) to obtain

$$\mathrm{KL}\left(N_y|N_x\right) = \frac{1}{2}\sum_{i=1}^{q}\log\hat{a}_i + \frac{N-q}{2}\log\left(\frac{1}{N-q}\sum_{i=q+1}^{N}\hat{a}_i\right) - \frac{1}{2}\sum_{j=1}^{N}\log\left(\lambda_j + \tau^{-1}\right). \tag{10}$$

Since we consider the matrix derived from the data $\mathbf{K}_y$ to be fixed, its trace stays constant. We write this constant as

$$\sum_{i=1}^{N}\hat{a}_i = \mathrm{tr}\left(\mathbf{K}_y\right) = \xi.$$

To minimise the KL divergence we will first upper bound it. Then we will show that the minimum of this bound conincides with the minimum of the KL divergence.

2.1.5 An Upper Bound on the Kullback Leibler Divergence

Jensen's inequality implies that

$$\sum_{i=1}^{q} \log \hat{a}_i \leq q \log \left[\frac{1}{q} \sum_{i=1}^{q} \hat{a}_i \right].$$

Substituting into (10) we obtain an upper bound on the KL divergence

$$\begin{aligned} \text{KL } (N_x|N_y) &\leq \frac{1}{2} q \log \left[\frac{1}{q} \sum_{i=1}^{q} \hat{a}_i \right] + \frac{N-q}{2} \log \left(\frac{1}{N-q} \sum_{i=q+1}^{N} \hat{a}_i \right) - \frac{1}{2} \left[\sum_{i=1}^{N} \log \left(\lambda_i + \tau^{-1} \right) \right] \\ &= \frac{1}{2} \log \frac{(\xi - x)^q \, x^{N-q}}{q^q \, (N-q)^{N-q}} - \frac{1}{2} \left[\sum_{i=1}^{N} \log \left(\lambda_i + \tau^{-1} \right) \right] \end{aligned}$$

where we have defined $x = \sum_{i=q+1}^{N} \hat{a}_i$.

2.1.6 The Minimum of the Bound

To find the minimum of the bound, we will first show that x lies in a constrained range of values, we will then demonstrate that across this range the value of the bound is strictly increasing. This implies that to minimise the bound x must take on its smallest value.

The smallest possible value for x is obtained when $N-q$ eigenvalues of $\mathbf{K}'_y$ are 0, in which case $x = (N-q)\,\tau^{-1}$. The largest value of x is obtained when q eigenvalues of $\mathbf{K}'_y$ are 0 (corresponding to $\hat{a}_1, \ldots, \hat{a}_q$), giving $x = \left(\xi - q\tau^{-1}\right)$. So we have $(N-q)\,\tau^{-1} \leq x \leq \left(\xi - q\tau^{-1}\right)$.

The bound on the KL divergence will be minimized when

$$f(x) = (\xi - x)^q \, x^{N-q}$$

is minimal, as the logarithm is a monotonic function. The function $f(x)$ has a global maximum at $x = \xi \frac{N-q}{N} \geq \left(\xi - q\tau^{-1}\right)$, hence in the interval under consideration f is strictly increasing. The bound will hence be minimized when x assumes its minimum value (the explicit value depending on the form of the "data" matrix $\mathbf{K}_y$).

2.1.7 Geometric Interpretation of the Minimum

As we have seen in the previous section, the minimum of the bound is obtained when $x = \sum_{i=q+1}^{N} \hat{a}_i$ is minimal. If the eigenvectors of $\mathbf{K}_x$ are unconstrained, this means that the space spanned by them will align with the q principal vectors of $\mathbf{K}_y$, as happens in PCA. If the eigenvectors are constrained, equation 9 shows that they will align to the principal subspace as closely as the bounds allow.

2.1.8 Coincidence of the Minima for the Divergence and its Bound

We now show that at this point we also have a minimum of the KL divergence itself. We will proceed by computing the variation of the KL divergence about this point. We denote the minimum values of $\hat{a}_i$ with $\bar{a}_i$.

As we saw in the previous paragraph, the data space is split (orthogonally) into two subspaces, the maximal on spanned by the first q eigenvectors of $\mathbf{K}_x$, and the minimal one which is its orthogonal complement. We will call maximal vectors the vectors from the maximal subspace, and minimal ones the ones from the minimal subspace.

We want to prove that this splitting of the space corresponds to a minimum of the KL divergence (as well as of its upper bound). Consider adding a variation δa_j to the element $\hat{a}_j$, $j \leq q$. [2]

[2] As we are interested in proving that $\bar{a}_i$ is a local minimum for the objective function, we restrict our analysis to a small neighbourhood of $\bar{a}_i$. A transformation can be represented as a small 'infinitesimal' vector δa_i.

As, by hypothesis, we are changing the split of the space, this will amount to add an infinitesimal component in a minimal direction to (at least one) maximal vector. Such a transformation must necessarily induce a negative variation in $\hat{a}_j$; also, given the constancy of $\sum_{i=1}^{N} \hat{a}_i$, it must induce an opposite variation in $\sum_{i=q+1}^{N} \hat{a}_i$ (if we rotate the maximal subspace, its orthogonal complement must also rotate). Then, ignoring the constant term, we get

$$KL\left(N_y|N_x;\bar{a}_i+\delta a_j\right)=\frac{1}{2}\sum_{i=1}^{q}\log\left(\bar{a}_i+\delta a_j\right)+\left(\frac{N-q}{2}\right)\log\left[\frac{1}{N-q}\left(\sum_{i=q+1}^{N}\bar{a}_i-\delta a_j\right)\right]\simeq$$

$$KL\left(N_y|N_x;\bar{a}_i\right)+\left[\frac{1}{\bar{a}_j}-\frac{(N-q)}{\sum_{i=q+1}^{N}\bar{a}_i}\right]\delta a_j. \tag{11}$$

The variation term in 11 is the product of the difference between the inverse of $\bar{a}_j$ (which belongs to the maximal subspace) and inverse of the mean of the minimal $\bar{a}_i$ (hence a negative number) with the variation δa_j, which is itself negative. Therefore, $\bar{a}_i$ is a (local) minimum of the KL divergence.

3 Block diagonal approximating kernels

As an illustration of the results of the previous section, we now turn to the case when the approximating kernel is forced to be block diagonal. In this case the constraints are linear and the solution to the constrained optimisation problem can be explicitly found in terms of the eigendecomposition of $\mathbf{K}_y$.

For simplicity's sake, we assume that the approximating kernel $\mathbf{K}_x = \mathbf{K}'_x + \beta^{-1}\mathbf{I}$ is made up of two diagonal blocks, of size p and $N-p$ respectively, and that each block is modelled as a rank 1 matrix plus a spherical term. This is equivalent to assuming that the matrix of eigenvalues $\mathbf{D}$ in 2 has only two nonzero entries, which we can assume (without loss of generality) to be the first and the $(p+1)$-the entries.

From the previous section we know that minimising the KL divergence $\mathrm{KL}\left(N_y \| N_x\right)$ is equivalent to maximising the quantities

$$a_i = \sum_{j=1}^{N} \lambda_j \left(\mathbf{u}_j^T \mathbf{v}_i\right)^2 = \mathbf{v}_i^{\mathrm{T}} \mathbf{K}_y \mathbf{v}_i \tag{12}$$

where N is the dimensionality of the space, λ are the eigenvalues of $\mathbf{K}_y$ and $\mathbf{u}_j$ and $\mathbf{v}_i$ are the eigenvectors of $\mathbf{K}_y$ and $\mathbf{K}_x$ respectively.

The block diagonal constraint means that we need to constrain the $\mathbf{v}$s in 12 to lie in particular subspaces; from 12 it is obvious that the solution will be for each of the $\mathbf{v}$s to align with the principal direction of each block.

3.1 Constrained principal components vs unconstrained principal components

It is sometimes the case though that the approximately block diagonal structure is not known in advance, and therefore it is more convenient to consider the full eigendecomposition of the kernel $\mathbf{K}_y$ rather than the decomposition of the blocks. This is the case for instance in clustering applications.

It is therefore interesting to compare the principal direction of a block with the restriction to the block of the principal direction of the whole matrix $\mathbf{K}_y$. We assume the kernel $\mathbf{K}_y$ to be only approximately diagonal, and view this as a perturbed version of an exactly block diagonal kernel $\hat{\mathbf{K}}_y$.

Consider the left uppermost block, of size q, and constrain the corresponding approximating $\mathbf{v}$ to be nonzero only in the first q components. Imagine for simplicity that the unperturbed matrix $\hat{\mathbf{K}}_y$ had only two non-zero eigenvalues in the upper block, say $\lambda_1 > \lambda_2$, with respective eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$. Then the unperturbed constrained solution would be $\mathbf{v} = \mathbf{u}_1$. Suppose now that $\mathbf{u}_1$ is perturbed adding components in the last $N-q$ directions, so that $\mathbf{u}_1 = \mathbf{u}_1^q + \mathbf{u}_1^{N-q}$, still maintaining $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_2 = 0$. The question is, how do we change $\mathbf{v}$ to keep this into account?

Let's start by writing $\mathbf{v} = p_1 \frac{\mathbf{u}_1^q}{|\mathbf{u}_1^q|} + p_2\mathbf{u}_2$ (obviously $\mathbf{v}$ will not acquire any component along the eigenvectors with eigenvalue zero, as this would not contribute to the sum in 12). We can rewrite our objective function in terms of $p_{1,2}$ as

$$a = \lambda_1 p_1^2 \left|\mathbf{u}_1^q\right|^2 + \lambda_2 p_2^2 \left|\mathbf{u}_2\right|^2 = \lambda_1 p_1^2 \left|\mathbf{u}_1^q\right|^2 + \lambda_2 p_2^2$$

as $\mathbf{u}_2$ is still a normalised eigenvector. Now we impose normalisation of $\mathbf{v}$ which implies

$$p_1^2 + p_2^2 = 1. \tag{13}$$

Defining $z = p_1^2$ we can rewrite the last two equations as

$$\begin{aligned} p_2^2 &= 1 - z \\ a &= z\left(\lambda_1 \left|\mathbf{u}_1^q\right|^2 - \lambda_2\right) + \lambda_2. \end{aligned}$$

Therefore, as a is linear in z, its maximum is readily found and depends on the sign of the coefficient $\alpha = \lambda_1 \left|\mathbf{u}_1^q\right|^2 - \lambda_2$; if $\alpha > 0$,then the maximum is obtained for the maximum of z at $z = p_1^2 = 1$; if $\alpha < 0$, the maximum is obtained for $z = 0$ and $p_2 = 1$. Hence the behaviour of the maximum is discontinuous: it's aligned with the (truncated) highest eigenvector for perturbations up to a certain value, and then switches instantly in line with the other eigenvector (thus changing to an orthogonal direction). This is easily extended to more than one eigenvector and more than one approximating direction.

The above discussion suggests that the true minimum of the KL divergence, being given by the eigenvector of the block, keeps into account both the eigenvalue associated with a direction and the entries of the eigenvector: specifically, it will seek to align to the direction with maximum eigenvalue which also retains most of its mass in the block.

4 Discussion

In this paper we considered a general framework for the optimisation of the KL divergence between two zero mean Gaussian distributions (represented by the associated kernel). We proved that the optimisation problem can be recast in term of a simpler problem involving the eigenvectors of the approximating kernel. When these are unconstrained, the minimisation returns standard probabilistic PCA (or Kernel PCA in the case of non-linear kernels). However, the simplified problem is manageble even in the case when the eigenvectors are constrained.

As an example, we discussed the a constraint in which the approximating kernel is block diagonal. The behaviour is relatively straightforward and allows for interesting connections with other unsupervised kernel techniques such as spectral clustering (see *e.g.* [4]). In particular, a comparison between the exact block diagonal solution and the one obtained by truncating the principal components of the kernel reveals some interesting properties which could give a principled explanation of widely used techniques such as deflation in image segmentation. A full discussion of the implications of these considerations, as well as of the suitability of KL divergence as an objective function in spectral clustering, is however beyond the scope of this paper.

References

[1] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[2] Neil D. Lawrence. Gaussian process models for visualisation of high dimensional data. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, pages 329–336, Cambridge, MA, 2004. MIT Press.

[3] Neil D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. Technical Report CS-04-08, Department of Computer Science, University of Sheffield, 2004.

[4] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In Thomas G. Dietterich, Sue Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, Cambridge, MA, 2002. MIT Press.

[5] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6(3):611–622, 1999.