Variational Principal Components

Christopher M. Bishop

Microsoft Research
7 J. J. Thomson Avenue, Cambridge, CB3 0FB, U.K.
cmbishop@microsoft.com
http://research.microsoft.com/∼cmbishop

In *Proceedings Ninth International Conference on Artificial Neural Networks*, ICANN'99, IEE, volume **1**, pages 509–514.

Abstract

One of the central issues in the use of principal component analysis (PCA) for data modelling is that of choosing the appropriate number of retained components. This problem was recently addressed through the formulation of a Bayesian treatment of PCA (Bishop, 1999a) in terms of a probabilistic latent variable model. A central feature of this approach is that the effective dimensionality of the latent space (equivalent to the number of retained principal components) is determined *automatically* as part of the Bayesian inference procedure. In common with most non-trivial Bayesian models, however, the required marginalizations are analytically intractable, and so an approximation scheme based on a local Gaussian representation of the posterior distribution was employed. In this paper we develop an alternative, variational formulation of Bayesian PCA, based on a factorial representation of the posterior distribution. This approach is computationally efficient, and unlike other approximation schemes, it maximizes a rigorous lower bound on the marginal log probability of the observed data.

1 Introduction

Principal component analysis (PCA) is a widely used technique for data analysis (Jolliffe, 1986). Recently Tipping and Bishop (1999b) showed that a specific form of generative latent variable model has the property that its maximum likelihood solution extracts the principal sub-space of the observed data set. This probabilistic reformulation of PCA permits many extensions including a principled formulation of mixtures of principal component analyzers, as discussed by Tipping and Bishop (1997, 1999a).

A central issue in maximum likelihood (as well as conventional) PCA is the choice of the number of principal components to be retained. This is particularly problematic in a mixture modelling context since ideally we would like the components to have potentially different dimensionalities. However, an exhaustive search over the choice of dimensionality for each of the components in a mixture distribution is often computationally intractable. In this paper we develop a Bayesian treatment of PCA based on variational inference, and we show how this leads to an *automatic* selection of the appropriate model dimensionality. Our approach avoids a discrete model search, involving instead the use of continuous hyper-parameters to determine an *effective* number of principal components.

2 Probabilistic PCA

Consider a data set D of observed d-dimensional vectors $D = \{\mathbf{t}_n\}$ where $n \in \{1, \ldots, N\}$. Conventional principal component analysis is obtained by first computing the sample covariance matrix given by

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{t}_n - \overline{\mathbf{t}})(\mathbf{t}_n - \overline{\mathbf{t}})^{\mathrm{T}} \qquad (1)$$

where $\overline{\mathbf{t}} = N^{-1} \sum_n \mathbf{t}_n$ is the sample mean. Next the eigenvectors $\mathbf{u}_i$ and eigenvalues λ_i of $\mathbf{S}$ are found, where $\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ and $i = 1, \ldots, d$. The eigenvectors corresponding to the q largest eigenvalues (where $q < d$) are retained, and a reduced-dimensionality representation of the data set is defined by $\mathbf{x}_n = \mathbf{U}_q^{\mathrm{T}}(\mathbf{t}_n - \overline{\mathbf{t}})$ where $\mathbf{U}_q = (\mathbf{u}_1, \ldots, \mathbf{u}_q)$. It is easily shown that PCA corresponds to the linear projection of a data set under which the retained variance is a maximum, or equivalently the linear projection for which the sum-of-squares reconstruction cost is minimized.

A significant limitation of conventional PCA is that it does not define a probability distribution. Recently, however, Tipping and Bishop (1999b) showed how PCA can be reformulated

as the maximum likelihood solution of a specific latent variable model, as follows. We first introduce a q-dimensional latent variable $\mathbf{x}$ whose prior distribution is a zero mean Gaussian

$$P(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I}_q) \tag{2}$$

where $\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{\Sigma})$ denotes a multivariate normal distribution over $\mathbf{x}$ with mean $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$, and $\mathbf{I}_q$ is the q-dimensional unit matrix. The observed variable $\mathbf{t}$ is then defined as a linear transformation of $\mathbf{x}$ with additive Gaussian noise $\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$ where $\mathbf{W}$ is a $d \times q$ matrix, $\boldsymbol{\mu}$ is a d-dimensional vector and $\boldsymbol{\epsilon}$ is a zero-mean Gaussian-distributed vector with covariance $\sigma^2\mathbf{I}_d$. Thus

$$P(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I}_d). \tag{3}$$

The marginal distribution of the observed variable is then given by the convolution of two Gaussians and is itself Gaussian

$$P(\mathbf{t}) = \int P(\mathbf{t}|\mathbf{x})P(\mathbf{x})\, d\mathbf{x} = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) \tag{4}$$

where the covariance matrix $\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}_d$. The model (4) represents a constrained Gaussian distribution governed by the parameters $\boldsymbol{\mu}$, $\mathbf{W}$ and σ^2.

The log probability of the parameters under the observed data set D is then given by

$$L(\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = -\frac{N}{2}\left\{ d\ln(2\pi) + \ln|\mathbf{C}| + \mathrm{Tr}\left[\mathbf{C}^{-1}\mathbf{S}\right] \right\} \tag{5}$$

where $\mathbf{S}$ is the sample covariance matrix given by (1). The maximum likelihood solution for $\boldsymbol{\mu}$ is easily seen to be $\boldsymbol{\mu}_{\mathrm{ML}} = \overline{\mathbf{t}}$. It was shown by Tipping and Bishop (1999b) that the stationary points of the log likelihood with respect to $\mathbf{W}$ satisfy

$$\mathbf{W}_{\mathrm{ML}} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I}_q)^{1/2} \tag{6}$$

where the columns of $\mathbf{U}_q$ are eigenvectors of $\mathbf{S}$, with corresponding eigenvalues in the diagonal matrix $\mathbf{\Lambda}_q$. It was also shown that the *maximum* of the likelihood is achieved when the q largest eigenvalues are chosen, so that the columns of $\mathbf{U}_q$ correspond to the *principal* eigenvectors, with all other choices of eigenvalues corresponding to saddle points. The maximum likelihood solution for σ^2 is then given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{d-q}\sum_{i=q+1}^{d} \lambda_i$$

which has a natural interpretation as the average variance lost per discarded dimension. The density model (4) thus represents a probabilistic formulation of PCA. It is easily verified that conventional PCA is recovered in the limit $\sigma^2 \to 0$. Some aspects of the links between latent variable models and PCA were noted independently by Roweis and Ghahramani (1999).

Probabilistic PCA has been successfully applied to problems in data compression, density estimation and data visualization (Tipping and Bishop, 1999a), and has been extended to mixture and hierarchical mixture models (Bishop and Tipping, 1998). As with conventional PCA, however, the model itself provides no mechanism for determining the value of the latent-space dimensionality q. For $q = d-1$ the model is equivalent to a full-covariance Gaussian distribution, while for $q < d-1$ it represents a constrained Gaussian in which the variance is independent in q directions, and in the remaining $d-q$ directions is governed by the single parameter σ^2. Thus the choice of q corresponds to a problem in model complexity optimization. If data is plentiful, then cross-validation to compare all possible values of q offers a possible, if computationally demanding, approach. However, for mixtures of probabilistic PCA models it quickly becomes intractable to explore the exponentially many combinations of different q values for each component. This problem can be resolved through an appropriate Bayesian reformulation.

3 Bayesian PCA

Armed with the probabilistic reformulation of PCA defined in Section 2, a Bayesian treatment of PCA is obtained by first introducing a prior distribution $P(\boldsymbol{\mu}, \mathbf{W}, \sigma^2)$ over the parameters of the model. The corresponding posterior distribution $P(\boldsymbol{\mu}, \mathbf{W}, \sigma^2|D)$ is then obtained by multiplying the prior by the likelihood function, whose logarithm is given by (5), and normalizing. Finally, the predictive density is obtained by marginalizing over the parameters, so that

$$P(\mathbf{t}|D) = \iiint P(\mathbf{t}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) P(\boldsymbol{\mu}, \mathbf{W}, \sigma^2|D)\, d\boldsymbol{\mu}\, d\mathbf{W}\, d\sigma^2.$$

In order to implement this framework we must address two issues: (i) the choice of prior distribution, and (ii) the formulation of a tractable procedure for marginalization over the posterior distribution. Our focus in this paper is on the specific issue of controlling the effective dimensionality of the latent space (correspond-

ing to the number of retained principal components). Furthermore, we seek to avoid discrete model selection and instead use continuous hyper-parameters to determine automatically the appropriate dimensionality for the latent space as part of the process of Bayesian inference. This is achieved by introducing a *hierarchical* prior $P(\mathbf{W}|\boldsymbol{\alpha})$ over the matrix $\mathbf{W}$, governed by a q-dimensional vector of hyper-parameters $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_q\}$. Each hyper-parameter controls one of the columns of the matrix $\mathbf{W}$ through a conditional Gaussian distribution of the form

$$P(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{i=1}^{q} \left(\frac{\alpha_i}{2\pi}\right)^{d/2} \exp\left\{-\frac{1}{2}\alpha_i \|\mathbf{w}_i\|^2\right\}$$

where $\{\mathbf{w}_i\}$ are the columns of $\mathbf{W}$. This form of prior is motivated by the framework of *automatic relevance determination* (ARD) introduced in the context of neural networks by Neal and MacKay (see MacKay, 1995). Each α_i controls the inverse variance of the corresponding $\mathbf{w}_i$, so that if a particular α_i has a posterior distribution concentrated at large values, the corresponding $\mathbf{w}_i$ will tend to be small, and that direction in latent space will be effectively 'switched off'. The dimensionality of the latent space is set to its maximum possible value $q = d - 1$.

The probabilistic structure of the model is displayed graphically in Figure 1.

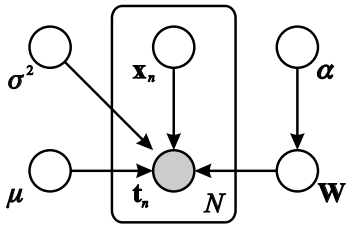

Figure 1: Representation of Bayesian PCA as a probabilistic graphical model showing the hierarchical prior over $\mathbf{W}$ governed by the vector of hyper-parameters $\boldsymbol{\alpha}$. The box denotes a 'plate' comprising a data set of N independent observations of the visible vector $\mathbf{t}_n$ (shown shaded) together with the corresponding latent variables $\mathbf{x}_n$.

We can complete the specification of the Bayesian model by defining (broad) priors over the parameters $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$ and σ^2. Specifically, defining $\tau \equiv \sigma^{-2}$, we make the following choices

$$\begin{aligned} P(\boldsymbol{\mu}) &= \mathcal{N}(\boldsymbol{\mu}|\mathbf{0}, \beta^{-1}\mathbf{I}) \\ P(\boldsymbol{\alpha}) &= \prod_{i=1}^{q} \Gamma(\alpha_i|a_\alpha, b_\alpha) \\ P(\tau) &= \Gamma(\tau|c_\tau, d_\tau). \end{aligned}$$

where $\Gamma(x|a, b)$ denotes a Gamma distribution over x given by

$$\Gamma(x|a, b) = \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)} \tag{7}$$

and $\Gamma(a)$ is the Gamma function. The distribution (7) has the useful properties

$$\langle x \rangle = \frac{a}{b} \tag{8}$$

$$\langle x^2 \rangle - \langle x \rangle^2 = \frac{a}{b^2}. \tag{9}$$

We obtain broad priors by setting $a_\alpha = b_\alpha = a_\tau = b_\tau = 10^{-3}$ and $\beta = 10^{-3}$.

In order to make use of this model in practice we must be able to marginalize with respect to $\mathbf{W}$, $\boldsymbol{\alpha}$, $\boldsymbol{\mu}$ and σ^2, as well as the latent variables $\{\mathbf{x}_n\}$. In Bishop (1999a) we considered two approximations based on type-II maximum likelihood using a local Gaussian approximation to the posterior distribution, and Markov chain Monte Carlo based on Gibbs sampling. In this paper we develop a variational treatment which is computationally efficient and which optimizes a rigorous bound on the marginal log probability.

4 Variational Inference

In order to motivate the variational approach, consider the problem of evaluating the marginal likelihood

$$P(D) = \int P(D, \boldsymbol{\theta})\, d\boldsymbol{\theta}$$

where $\boldsymbol{\theta} = \{\theta_i\}$ denotes the set of all parameters and latent variables in the model. We have already noted that such integrations are analytically intractable. Variational methods involve the introduction of a distribution $Q(\boldsymbol{\theta})$ which, as we shall see shortly, provides an approximation to the true posterior distribution. Consider the following transformation applied to the log marginal likelihood

$$\begin{aligned} \ln P(D) &= \ln \int P(D, \boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \ln \int Q(\boldsymbol{\theta}) \frac{P(D, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\ &\geq \int Q(\boldsymbol{\theta}) \ln \frac{P(D, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\ &= \mathcal{L}(Q) \end{aligned} \tag{10}$$

where we have applied Jensen's inequality. We see that the function $\mathcal{L}(Q)$ forms a rigorous lower bound on the true log marginal likelihood.

The significance of this transformation is that, through a suitable choice for the Q distribution, the quantity $\mathcal{L}(Q)$ may be tractable to compute, even though the original log likelihood function is not. From (10) it is easy to see that the difference between the true log marginal likelihood $\ln P(D)$ and the bound $\mathcal{L}(Q)$ is given by

$$\mathrm{KL}(Q\|P) = -\int Q(\boldsymbol{\theta}) \ln \frac{P(\boldsymbol{\theta}|D)}{Q(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \quad (11)$$

which is the Kullback-Leibler (KL) divergence between the approximating distribution $Q(\boldsymbol{\theta})$ and the true posterior $P(\boldsymbol{\theta}|D)$. The relationship between the various quantities is shown in Figure 2.

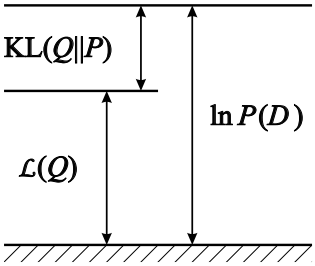

Figure 2: The quantity $\mathcal{L}(Q)$ provides a rigorous lower bound on the true log marginal likelihood $\ln P(D)$, with the difference being given by the Kullback-Leibler divergence $\mathrm{KL}(Q\|P)$ between the approximating distribution $Q(\boldsymbol{\theta})$ and the true posterior $P(\boldsymbol{\theta}|D)$.

The goal in a variational approach is to choose a suitable form for $Q(\boldsymbol{\theta})$ which is sufficiently simple that the lower bound $\mathcal{L}(Q)$ can readily be evaluated and yet which is sufficiently flexible that the bound is reasonably tight. We generally choose some family of Q distributions and then seek the best approximation within this family by maximizing the lower bound. Since the true log likelihood is independent of Q we see that this is equivalent to minimizing the Kullback-Leibler divergence.

Suppose we consider a completely free-form optimization over Q, allowing for all possible Q distributions. Using the well-known result that the KL divergence between two distributions $Q(\boldsymbol{\theta})$ and $P(\boldsymbol{\theta})$ is minimized by $Q(\boldsymbol{\theta}) = P(\boldsymbol{\theta})$ we see that the optimal Q distribution is given by the true posterior, in which case the KL divergence is zero and the bound becomes exact. However, this will not lead to any simplification of the problem. In order to make progress it is necessary to consider a more restricted range of Q distributions.

One approach is to consider a parametric family of Q distributions of the form $Q(\boldsymbol{\theta}, \boldsymbol{\chi})$ governed by a set of parameters $\boldsymbol{\chi}$. We can then adapt $\boldsymbol{\chi}$ by minimizing the KL divergence to find the best approximation within this family. Here we consider an alternative approach which is to restrict the functional form of $Q(\boldsymbol{\theta})$ by assuming that it factorizes over the component variables $\{\theta_i\}$ in $\boldsymbol{\theta}$, so that

$$Q(\boldsymbol{\theta}) = \prod_i Q_i(\theta_i). \quad (12)$$

The KL divergence can then be minimized over all possible factorial distributions by performing a free-form minimization over the Q_i, leading to the following result

$$Q_i(\theta_i) = \frac{\exp\left\langle \ln P(D, \boldsymbol{\theta})\right\rangle_{k\neq i}}{\int \exp\left\langle \ln P(D, \boldsymbol{\theta})\right\rangle_{k\neq j} d\theta_j} \quad (13)$$

where $\langle\ \cdot\ \rangle_{k\neq i}$ denotes an expectation with respect to the distributions $Q_k(\theta_k)$ for all $k \neq i$.

In order to apply this framework to Bayesian PCA we assume a Q distribution of the form

$$\begin{aligned} &Q(X, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \tau) \\ &\quad = \quad Q(X)Q(\mathbf{W})Q(\boldsymbol{\alpha})Q(\boldsymbol{\mu})Q(\tau) \quad (14) \end{aligned}$$

where $X = \{\mathbf{x}_n\}$. The true joint distribution of data and parameters is given by

$$\begin{aligned} P(D, \theta) = \prod_{n=1}^{N} P(\mathbf{t}_n|\mathbf{x}_n, \mathbf{W}, \boldsymbol{\mu}, \tau)P(X) \\ P(\mathbf{W}|\boldsymbol{\alpha})P(\boldsymbol{\alpha})P(\boldsymbol{\mu})P(\tau). \quad (15) \end{aligned}$$

Using the result (13), together with the explicit forms for the various $P(\cdot)$ distributions, we obtain the following results for the component distributions of $Q(\cdot)$

$$\begin{aligned} Q(X) &= \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\mathbf{m}_{\mathbf{x}}^{(n)}, \boldsymbol{\Sigma}_{\mathbf{x}}) \quad (16) \\ Q(\boldsymbol{\mu}) &= \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}) \quad (17) \\ Q(\mathbf{W}) &= \prod_{k=1}^{d} \mathcal{N}(\widetilde{\mathbf{w}}_k|\mathbf{m}_{\mathbf{w}}^{(k)}, \boldsymbol{\Sigma}_{\mathbf{w}}) \quad (18) \\ Q(\boldsymbol{\alpha}) &= \prod_{i=1}^{q} \Gamma(\alpha_i|\widetilde{a}_\alpha, \widetilde{b}_{\alpha i}) \quad (19) \\ Q(\tau) &= \Gamma(\tau|\widetilde{a}_\tau, \widetilde{b}_\tau) \quad (20) \end{aligned}$$

where $\widetilde{\mathbf{w}}_k$ denotes a column vector correspond-

ing to the kth row of $\mathbf{W}$, we have defined

$$\begin{aligned}
\mathbf{m}_{\mathbf{x}}^{(n)} &= \langle\tau\rangle\boldsymbol{\Sigma}_{\mathbf{x}}\langle\mathbf{W}^{\mathrm{T}}\rangle(\mathbf{t}_n - \langle\boldsymbol{\mu}\rangle) \\
\boldsymbol{\Sigma}_{\mathbf{x}} &= \left(\mathbf{I} + \langle\tau\rangle\langle\mathbf{W}^{\mathrm{T}}\mathbf{W}\rangle\right)^{-1} \\
\mathbf{m}_{\boldsymbol{\mu}} &= \langle\tau\rangle\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\sum_{n=1}^{N}(\mathbf{t}_n - \langle\mathbf{W}\rangle\langle\mathbf{x}_n\rangle) \\
\boldsymbol{\Sigma}_{\boldsymbol{\mu}} &= (\beta + N\langle\tau\rangle)^{-1}\mathbf{I} \\
\mathbf{m}_{\mathbf{w}}^{(k)} &= \langle\tau\rangle\boldsymbol{\Sigma}_{\mathbf{w}}\sum_{n=1}^{N}\langle\mathbf{x}_n\rangle(t_{nk} - \langle\mu_k\rangle) \\
\boldsymbol{\Sigma}_{\mathbf{w}} &= \left(\mathrm{diag}\langle\boldsymbol{\alpha}\rangle + \langle\tau\rangle\sum_{n=1}^{N}\langle\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle\right)^{-1} \\
\widetilde{a}_\alpha &= a_\alpha + \frac{d}{2} \\
\widetilde{b}_{\alpha i} &= b_\alpha + \frac{\langle\|\mathbf{w}_i\|^2\rangle}{2} \\
\widetilde{a}_\tau &= a_\tau + \frac{Nd}{2} \\
\widetilde{b}_\tau &= b_\tau + \frac{1}{2}\sum_{n=1}^{N}\big\{\|\mathbf{t}_n\|^2 + \langle\|\boldsymbol{\mu}\|^2\rangle \\
&\quad + \mathrm{Tr}(\langle\mathbf{W}^{\mathrm{T}}\mathbf{W}\rangle\langle\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle) \\
&\quad + 2\langle\boldsymbol{\mu}^{\mathrm{T}}\rangle\langle\mathbf{W}\rangle\langle\mathbf{x}_n\rangle \\
&\quad - 2\mathbf{t}_n^{\mathrm{T}}\langle\mathbf{W}\rangle\langle\mathbf{x}_n\rangle - 2\mathbf{t}_n^{\mathrm{T}}\langle\boldsymbol{\mu}\rangle\big\}
\end{aligned}$$

and $\mathrm{diag}\langle\boldsymbol{\alpha}\rangle$ denotes a diagonal matrix whose diagonal elements are given by $\langle\alpha_i\rangle$.

An interesting point to note is that we automatically obtain some additional factorization, in $Q(X)$, $Q(\mathbf{W})$ and $Q(\boldsymbol{\alpha})$, that was not assumed in the original Q distribution (14).

The solution for the optimal factors in the $Q(\boldsymbol{\theta})$ distribution is, of course, an implicit one since each distribution depends on moments of the other distributions. We can find a solution numerically by starting with a suitable initial guess for the distributions and then cycling through the groups of variables in turn, re-estimating each distribution using the above results. Note that at each re-estimation step we need only moments of those variables contained in the corresponding Markov blanket, which can be obtained from Figure 1.

The required moments are easily evaluated using (8) together with the result that, for a Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\langle\mathbf{x}\rangle = \boldsymbol{\mu} \tag{21}$$

$$\langle\mathbf{x}\mathbf{x}^{\mathrm{T}}\rangle = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}. \tag{22}$$

In order to monitor the convergence of the variational optimization it is convenient to be able to evaluate the lower bound $\mathcal{L}(Q)$ on the marginal log likelihood, which is easily done. As well as monitoring the evolution of the bound during training, the derivatives of the bound with respect to the parameters of the Q distribution at the end of the optimization process can be evaluated numerically using central differences to confirm that they are indeed close to zero.

5 Illustration

In this section we consider a simple example to illustrate the capability of the model to determine the appropriate number of principal components. We generate 100 data points in $d = 10$ dimensions from a Gaussian distribution having standard deviations of $(5, 4, 3, 2)$ along four orthogonal directions and a standard deviation of 1 in the remaining five directions. The result of fitting a maximum likelihood PCA model is shown as a Hinton diagram in Figure 3, in which the elements of $\mathbf{W}$ are given by (6).

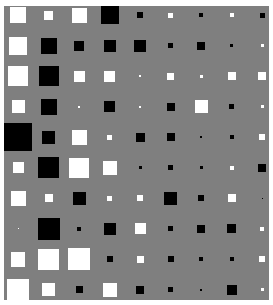

Figure 3: Hinton diagram of the elements of $\mathbf{W}$ in the maximum likelihood PCA model for a data set in 10 dimensions drawn from a distribution having different variances in 4 independent directions and a common, smaller variance in the remaining 6.

Each column of $\mathbf{W}$ corresponds to one potential principal direction in data space. Note that the eigenvector routine used to compute this solution finds orthogonal eigenvectors and orders them left to right in order of decreasing eigenvalue. We see that the first four columns have larger elements than the remaining columns. However, the last six columns have non-zero entries even though the data was generated from a distribution requiring only four principal components. In a finite data set the sample variance will be different along the different principal directions. The maximum likelihood solution cannot distinguish between signal and (sample) noise and so attributes each different variance value to the presence of an independent principal component.

Next we run variational Bayesian PCA on the same data set. The resulting matrix $\langle\mathbf{W}\rangle$ is shown in Figure 4. Here we see that the model

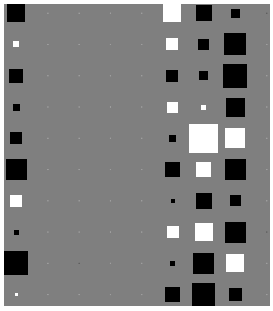

Figure 4: Hinton diagram of $\langle \mathbf{W} \rangle$ from variational Bayesian PCA for the same data set used to obtain Figure 3. Note that all but four of the columns of $\langle \mathbf{W} \rangle$ have been suppressed. Since the posterior variance of $\mathbf{W}$ is small, typical samples from this posterior have the same property.

has correctly discovered the underlying dimensionality of the generator of the data. Although the sample variance is different in the six low-variance directions there is insufficient evidence in the data set to support a complex model having more than four non-zero principal components.

6 Conclusions

An important application of probabilistic PCA models is to density modelling. Given the probabilistic formulation of PCA it is straightforward to construct a mixture distribution comprising a linear superposition of principal component analyzers. In the case of maximum likelihood PCA we have to choose both the number M of components and the latent space dimensionality q for each component. For moderate numbers of components and data spaces of several dimensions it quickly becomes intractable to explore the exponentially large number of combinations of q values for a given value of M. Here Bayesian PCA offers a significant advantage in allowing the effective dimensionalities of the models to be determined automatically. An example of mixture modelling with Bayesian PCA components (using the Laplace approximation) involving hand-written digit data is given in Bishop (1999a).

The Bayesian treatment of PCA discussed in this paper can be particularly advantageous for small data sets in high dimensions as it can avoid the singularities associated with maximum likelihood (or conventional) PCA by suppressing unwanted degrees of freedom in the model. This is especially helpful in a mixture modelling context, since the effective number of data points associated with specific 'clusters' can be small even when the total number of data points is large.

It should be emphasised that the Bayesian framework discussed in this paper does not determine a specific value for the number of non-zero principal components. Rather, it estimates a posterior distribution over models including those with a complete range of possible (effective) dimensionalities. For many applications this distribution may be tightly concentrated on a specific dimensionality. However, a posterior distribution over models is much more powerful than a point estimate of complexity. Optimal Bayesian predictions are obtained by marginalizing over all models, weighted by their posterior distribution.

Acknowledgements

I would like to thank Neil Lawrence for helpful discussions as well as for his contributions to the Matlab implementation of variational Bayesian PCA.

References

Bishop, C. M. (1999). Bayesian PCA. In S. A. S. M. S. Kearns and D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems*, Volume 11, pp. 382–388. MIT Press.

Bishop, C. M. and M. E. Tipping (1998). A hierarchical latent variable model for data visualization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (3), 281–293.

Jolliffe, I. T. (1986). *Principal Component Analysis*. New York: Springer-Verlag.

Roweis, S. and Z. Ghahramani (1999). A unifying review of linear Gaussian models. *Neural Computation* **11**, 305–345.

Tipping, M. E. and C. M. Bishop (1997). Mixtures of principal component analysers. In *Proceedings IEE Fifth International Conference on Artificial Neural Networks, Cambridge, U.K., July.*, pp. 13–18. London: IEE.

Tipping, M. E. and C. M. Bishop (1999a). Mixtures of probabilistic principal component analyzers. *Neural Computation* **11** (2), 443–482.

Tipping, M. E. and C. M. Bishop (1999b). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B* **21** (3), 611–622.