Constrained Subspace Modelling

Jaco Vermaak
Cambridge University Engineering Department
Cambridge, CB2 1PZ, UK

Patrick Pérez
Microsoft Research
Cambridge, CB3 0FB, UK

Abstract

When performing subspace modelling of data using Principal Component Analysis (PCA) it may be desirable to constrain certain directions to be more meaningful in the context of the problem being investigated. This need arises due to the data often being approximately isotropic along the lesser principal components, making the choice of directions for these components more-or-less arbitrary. Furthermore, constraining may be imperative to ensure viable solutions in problems where the dimensionality of the data space is of the same order as the number of data points available. This paper adopts a Bayesian approach and augments the likelihood implied by Probabilistic Principal Component Analysis (PPCA) [14] with a prior designed to achieve the constraining effect. The subspace parameters are computed efficiently using the EM algorithm. The constrained modelling approach is illustrated on two pertinent problems, one from speech analysis, and one from computer vision.

1. Introduction

An important aspect of modelling and visualising high dimensional data involves the removal of redundancies by finding a lower dimensional subspace that best captures the data characteristics. The most popular method for achieving this is Principal Component Analysis (PCA), which finds the linear embedding subspace that maximises the variance of the projected data.

In [14] PCA is reformulated in a probabilistic framework, coining the term Probabilistic Principal Component Analysis (PPCA). PPCA essentially augments the linear mapping from the PCA space to the observed data space by assuming the observed data to be corrupted by isotropic Gaussian noise, and the PCA coefficients to follow an isotropic Gaussian distribution in the embedding subspace. In the limiting case where the observation noise variance goes to zero standard PCA is recovered.

Reformulating PCA in a probabilistic framework has been key in the development of a large number of interesting extensions to PCA. Most importantly, the PPCA likelihood facilitates a Bayesian analysis under a wide variety of prior beliefs about the model parameters. Notable extensions include PPCA Mixture Models [13] to approximate non-linear manifolds, and Bayesian PCA [1], where the PPCA likelihood is augmented with an Automatic Relevance Determination (ARD) prior [11] to automatically estimate the dimensionality of the embedding subspace. A mixture version of Bayesian PCA has also been defined and applied to the problem of non-linear image modelling [2]. Other examples that are not explicitly formulated in a probabilistic framework include Robust PCA [8] to deal with data outliers, and Sparse PCA [5] to encourage sparseness of the principal components.

In most applications of PCA the largest proportion of the variance is captured by the first few principal components, with the variance in the remaining components often being approximately isotropic. Thus the selection of principal directions for these remaining components is more-or-less arbitrary, and they can be selected to be in some sense more meaningful in the context of the problem being studied. Sparse PCA [5] is one such example where the principal components are encouraged to have as few non-zero components as possible.

Within a Bayesian setting the notion of meaningful principal components can be elegantly represented by a suitably constructed prior distribution. This paper introduces a prior that is a generalisation of the ARD prior used in Bayesian PCA in the sense that each element of the principle components is allowed to have its own mean and variance. As opposed to Bayesian PCA where the variances are considered as hyperparameters to be estimated alongside the PCA parameters, the means and variances are fixed to problem-dependent values, with the effect of constraining the resulting PCA subspace. The prior is conjugate to the PPCA likelihood function, so that MAP estimates of the model parameters can be obtained by the EM algorithm [7].

A prior of this form is not only useful to constrain the principal components to meaningful values in cases where the data is approximately isotropic, but also plays an impor-

1063-6919/03 $17.00 © 2003 IEEE

tant role in problems where the dimensionality of the data space is of the same order or higher than the number of data points available. Such data sets often arise in *e.g.* computer vision applications, and may lead to data covariance matrices that are rank deficient. In these cases a prior is not only desirable, but imperative.

This constrained subspace modelling approach is illustrated on two problems, one from speech analysis, and one from computer vision. The objective in the speech analysis problem is to construct a subspace model for speech spectra. Unconstrained PCA places too much emphasis on the low frequencies. However, constraining some of the principal components to focus on the mid and high frequencies leads to reconstructed signals with higher perceptual quality. In the computer vision problem constrained PCA is used to construct a subspace model for 2D frontal face images, where certain of the principal components are constrained to focus on certain key areas, such as the eyes and mouth. Such a model achieves essentially the same reconstruction performance as unconstrained PCA, but allows the deformations in the key areas to be captured by small subsets of the PCA coefficients.

The remainder of the paper is organised as follows. Section 2 presents the two components of the Constrained PCA Model, *i.e.* the PPCA likelihood and the constraining prior distribution. Section 3 derives an EM algorithm to compute MAP estimates of the parameters of this model. In Section 4 the constrained subspace modelling approach is illustrated on a speech analysis and a face modelling problem, and compared with the results obtained by unconstrained PCA. Finally, in Section 5 some conclusions are reached.

2. The Constrained PCA Model

The constrained PCA model is formulated in a Bayesian setting. This section presents the two components of the model. The PPCA likelihood has been derived before in [14], but is presented in Section 2.1 for the sake of completeness. Section 2.2 then introduces a prior distribution to achieve the desired constraining effect.

2.1. The PPCA Likelihood

In PPCA an observed data point $\mathbf{x} \in \mathbb{R}^d$ is related to a latent variable $\boldsymbol{\alpha} \in \mathbb{R}^k$, $k < d$, through the linear mapping

$$\mathbf{x} = \mathbf{B}\boldsymbol{\alpha} + \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

where $\mathbf{B} = [\mathbf{b}_1 \cdots \mathbf{b}_k] \in \mathbb{R}^{d\times k}$ is a matrix whose columns span the latent subspace, $\boldsymbol{\mu} \in \mathbb{R}^d$ is the data mean, and $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ is the observation noise. Both the latent variable and observation noise are assumed to follow zero mean isotropic Gaussian distributions of the form $p(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\alpha}|\mathbf{0}, \mathbf{I}_k)$ and $p(\boldsymbol{\varepsilon}) = \mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0}, \sigma^2\mathbf{I}_d)$, respectively, where $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. Conditional on the latent variable the observation likelihood is of the form

$$p(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\mathbf{B}\boldsymbol{\alpha} + \boldsymbol{\mu}, \sigma^2\mathbf{I}_d), \tag{1}$$

where $\boldsymbol{\theta} = (\mathbf{B}, \boldsymbol{\mu}, \sigma^2)$ denotes the unknown model parameters. By integrating this expression over the latent variable the marginal likelihood is obtained as

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\theta})p(\boldsymbol{\alpha})d\boldsymbol{\alpha} = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{B}\mathbf{B}^{\mathrm{T}} + \sigma^2\mathbf{I}_d). \tag{2}$$

For a data set $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_N] \in \mathbb{R}^{d\times N}$, with corresponding latent variables $A = [\boldsymbol{\alpha}_1 \cdots \boldsymbol{\alpha}_N] \in \mathbb{R}^{k\times N}$, the conditional and marginal likelihoods are simply the product of the expressions in (1) and (2), respectively, over all the data points, since these are assumed to be independent, yielding

$$p(\mathbf{X}|\mathbf{A}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i|\mathbf{B}\boldsymbol{\alpha}_i + \boldsymbol{\mu}, \sigma^2\mathbf{I}_d) \tag{3}$$

$$p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \mathbf{B}\mathbf{B}^{\mathrm{T}} + \sigma^2\mathbf{I}_d). \tag{4}$$

It is shown in [14] that the stationary points of the marginal likelihood in (4) with respect to $\mathbf{B}$ are given by

$$\mathbf{B}_{\mathrm{ML}} = \mathbf{V}(\boldsymbol{\Lambda} - \sigma^2\mathbf{I}_k)^{1/2}\mathbf{R},$$

where the columns of $\mathbf{V} \in \mathbb{R}^{d\times k}$ are eigenvectors of the data covariance matrix $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \in \mathbb{R}^{d\times d}$, with corresponding eigenvalues in the diagonal matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{k\times k}$, and $\mathbf{R} \in \mathbb{R}^{k\times k}$ is an arbitrary orthogonal rotation matrix. The maximum likelihood is achieved when the k largest eigenvalues are chosen, with all other choices of eigenvalues corresponding to saddle points. With $\mathbf{B} = \mathbf{B}_{\mathrm{ML}}$ the maximum likelihood estimate for the variance is given by

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

with λ_i the i-th principal eigenvalue. This estimate has the interpretation as the average variance lost per discarded dimension. Finally, the maximum likelihood estimate for the mean is simply the mean of the data, *i.e.* $\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$, since the observation noise is assumed to be zero mean Gaussian distributed. As shown in [14] the maximum likelihood parameters can also be computed using the EM algorithm. This avoids explicit computation of the data covariance matrix, and can lead to substantial computational savings if the dimensionality of the subspace is much lower than that of the original data space.

Under the modelling assumptions PPCA thus finds the subspace that maximises the projected data variance. This is equivalent to minimising the reconstruction error when the data is mapped back to the original data space. The relationship between PPCA and standard PCA becomes more evident when examining the posterior distribution of the latent variable, given by

$$p(\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\alpha}|\widehat{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}),$$

with

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} &= \sigma^2(\mathbf{B}^{\mathrm{T}}\mathbf{B} + \sigma^2\mathbf{I}_k)^{-1}\\ \widehat{\boldsymbol{\alpha}} &= \frac{1}{\sigma^2}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\mathbf{B}^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}).\end{aligned}$$

The optimal projection onto the latent space is given by the mean of this distribution. When the maximum likelihood subspace matrix $\mathbf{B}_{\mathrm{ML}}$ is used, and the variance σ^2 is forced to zero, it is straightforward to see that standard PCA is recovered.

2.2. The Constraining Prior Distribution

In unconstrained PCA the largest proportion of the variance is normally captured by the first few principal components, with the variance in the remaining components often being approximately isotropic. Thus the selection of principal directions for these remaining components is more-or-less arbitrary, and they can be chosen to be in some sense more meaningful in the context of the problem being investigated. Furthermore, in sparse data problems where the number of data points available is of the same order as the dimensionality of the data space, the PCA solution may be undefined, and it is desirable to constrain the subspace in some way.

Here this constrained modelling is achieved by augmenting the PPCA likelihood model with a prior distribution over the model parameters of the form

$$p(\boldsymbol{\theta}) = p(\mathbf{B})p(\boldsymbol{\mu})p(\sigma^2), \tag{5}$$

with

$$\begin{aligned}p(\mathbf{B}) &= \prod_{i=1}^{k} p(\mathbf{b}_i) = \prod_{i=1}^{k}\mathcal{N}(\mathbf{b}_i|\widehat{\mathbf{b}}_i, \boldsymbol{\Sigma}_{\mathbf{b}_i})\\ p(\boldsymbol{\mu}) &= \mathcal{N}(\boldsymbol{\mu}|\widehat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}})\\ p(\sigma^2) &= \mathcal{IG}(\sigma^2|a,b),\end{aligned}$$

where $\boldsymbol{\Sigma}_{\mathbf{b}_i} = \mathrm{diag}(\sigma^2_{\mathbf{b}_i,1}\cdots\sigma^2_{\mathbf{b}_i,d})$, and $\mathcal{IG}(x|a,b) \propto x^{-(a+1)}\exp(-b/x)$ denotes the inverted gamma distribution with parameters a and b. In the above the parameters $(\{\widehat{\mathbf{b}}_i, \boldsymbol{\Sigma}_{\mathbf{b}_i} : i = 1\cdots k\}, \widehat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}, a, b)$ are assumed to be fixed and known.

Each element of the principal components has its own associated prior mean and variance. The constraining effect is achieved by setting the means to reflect certain desired directions, and the variance to capture the desired preciseness of these directions. Typically only a subset of components are constrained. In the limit where all the subspace component variances approach infinity, and the inverted gamma prior on σ^2 is set to be uniform ($a = -1$ and $b = 0$), PPCA is recovered.

This prior stands in contrast with the one introduced in BPCA [1]. The latter is motivated by the framework of ARD [11], and is given by

$$p(\mathbf{B}) = \prod_{i=1}^{k} p(\mathbf{b}_i) = \prod_{i=1}^{k}\mathcal{N}(\mathbf{b}_i|\mathbf{0}, \sigma^2_{\mathbf{b}_i}\mathbf{I}_d).$$

Thus a single variance is associated with each zero mean principal component. The variances are assumed to be inverted gamma distributed hyperparameters, and are estimated alongside the other parameters. Components with small posterior variances are effectively switched off. Thus the objective of BPCA is to automatically determine the effective dimensionality of the latent subspace, and not to constrain it in any other way, which is the objective here. It is, however, possible to incorporate the effect of BPCA in our model by associating an additional variance component with each of the subspace vectors, and estimating these subject to an ARD prior.

3. An EM Estimation Algorithm

The objective here is to find the MAP estimate of the model parameters, defined as

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}),$$

with the marginal posterior given by $p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Direct maximisation of the marginal posterior is analytically intractable, but a solution may be obtained iteratively by the EM algorithm [7], which alternates over the steps

$$\begin{aligned}&\textbf{E-step:} \quad Q(\boldsymbol{\theta},\boldsymbol{\theta}') = \langle \log p(\boldsymbol{\theta},\mathbf{A}|\mathbf{X})\rangle_{p(\mathbf{A}|\mathbf{X},\boldsymbol{\theta}')}\\ &\textbf{M-step:} \quad \boldsymbol{\theta}' \leftarrow \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}'),\end{aligned}$$

from some initial estimate, until convergence is achieved. In the above $\langle\cdot\rangle_p$ denotes the expectation operator relative to the distribution p, and the full posterior and the conditional posterior for the latent variables are given by

$$\begin{aligned}p(\boldsymbol{\theta},\mathbf{A}|\mathbf{X}) &\propto p(\mathbf{X}|\mathbf{A},\boldsymbol{\theta})p(\mathbf{A})p(\boldsymbol{\theta})\\ p(\mathbf{A}|\mathbf{X},\boldsymbol{\theta}) &= \prod_{i=1}^{N} p(\boldsymbol{\alpha}_i|\mathbf{x}_i,\boldsymbol{\theta}) = \prod_{i=1}^{N}\mathcal{N}(\boldsymbol{\alpha}_i|\widehat{\boldsymbol{\alpha}}_i, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}),\end{aligned}$$

where $p(\mathbf{X}|\mathbf{A},\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are given by (3) and (5), respectively, and $p(\mathbf{A})=\prod_{i=1}^{N}p(\boldsymbol{\alpha}_i)$. Substituting these expressions into the E-step above and simplifying, lead to an expression for the Q-function, given by

$$\begin{aligned}Q(\boldsymbol{\theta},\boldsymbol{\theta}') = -\frac{1}{2\sigma^2}&[\mathrm{tr}[(\mathbf{X}{-}\mathbf{U})(\mathbf{X}{-}\mathbf{U})^{\mathrm{T}}{-}2(\mathbf{X}{-}\mathbf{U})\langle\mathbf{A}\rangle^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\\ &+\mathbf{B}(N\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}+\langle\mathbf{A}\rangle\langle\mathbf{A}\rangle^{\mathrm{T}})\mathbf{B}^{\mathrm{T}}]+2b]\\ &-\frac{1}{2}\mathrm{vec}(\mathbf{B}-\widehat{\mathbf{B}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{B}}^{-1}\mathrm{vec}(\mathbf{B}-\widehat{\mathbf{B}})\\ -\frac{1}{2}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})&-\frac{1}{2}[Nd+2(a+1)]\log\sigma^2,\end{aligned}$$

where terms independent of the model parameters have been discarded. In the above $\langle\mathbf{A}\rangle = [\langle\boldsymbol{\alpha}_1\rangle\cdots\langle\boldsymbol{\alpha}_N\rangle]\in\mathbb{R}^{k\times N}$, $\widehat{\mathbf{B}}=[\widehat{\mathbf{b}}_1\cdots\widehat{\mathbf{b}}_k]\in\mathbb{R}^{d\times k}$, $\boldsymbol{\Sigma}_{\mathbf{B}}=\mathrm{diag}(\boldsymbol{\Sigma}_{\mathbf{b}_1}\cdots\boldsymbol{\Sigma}_{\mathbf{b}_k})\in\mathbb{R}^{dk\times dk}$, $\mathbf{U}=\boldsymbol{\mu}\otimes\mathbf{1}_{1\times N}\in\mathbb{R}^{d\times N}$, with $\mathbf{1}_{n\times m}$ denoting the $n\times m$ matrix of ones. The operator $\otimes$ denotes the Kronecker matrix product, tr$(\cdot)$, the matrix trace operator, and vec$(\cdot)$, the operator that forms a vector by stacking the columns of its matrix argument.

The E-step essentially entails the computation of the sufficient statistics for the distribution $p(\mathbf{A}|\mathbf{X},\boldsymbol{\theta})$. Since this is a product of Gaussian distributions with a common covariance matrix, the sufficient statistics follow straightforwardly as

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}&=\sigma^2(\mathbf{B}^{\mathrm{T}}\mathbf{B}+\sigma^2\mathbf{I}_k)^{-1}\\ \langle\mathbf{A}\rangle&=\frac{1}{\sigma^2}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\mathbf{B}^{\mathrm{T}}(\mathbf{X}-\mathbf{U}).\end{aligned}$$

Conditional on these values the M-step then maximises the Q-function with respect to the model parameters. Setting the derivatives of the Q-function with respect to the model parameters to zero leads, after some algebraic manipulation, to expressions for the stationary points of the Q-function, given by

$$\begin{aligned}&\mathrm{vec}(\mathbf{B})=[(N\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}+\langle\mathbf{A}\rangle\langle\mathbf{A}\rangle^{\mathrm{T}})\otimes\mathbf{I}_d+\sigma^2\boldsymbol{\Sigma}_{\mathbf{B}}^{-1}]^{-1}\\ &\quad\times[\mathrm{vec}((\mathbf{X}-\mathbf{U})\langle\mathbf{A}\rangle^{\mathrm{T}})+\sigma^2\boldsymbol{\Sigma}_{\mathbf{B}}^{-1}\mathrm{vec}(\widehat{\mathbf{B}})]\\ &\boldsymbol{\mu}=(N\mathbf{I}_d+\sigma^2\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1})^{-1}[(\mathbf{X}-\mathbf{B}\langle\mathbf{A}\rangle)\mathbf{1}_{N\times 1}+\sigma^2\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\widehat{\boldsymbol{\mu}}]\\ &\sigma^2=\frac{1}{Nd+2(a+1)}[2b+\mathrm{tr}[(\mathbf{X}-\mathbf{U})(\mathbf{X}-\mathbf{U})^{\mathrm{T}}\\ &\quad-2(\mathbf{X}-\mathbf{U})\langle\mathbf{A}\rangle^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}+\mathbf{B}(N\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}+\langle\mathbf{A}\rangle\langle\mathbf{A}\rangle^{\mathrm{T}})\mathbf{B}^{\mathrm{T}}]].\end{aligned}$$

These expressions are coupled, but can be solved by cycling over them, solving each in turn while fixing the values for the other parameters to their current estimates, until convergence is achieved. Due to the locally quadratic approximation such a procedure normally converges within a small number of iterations.

The computation of the sufficient statistics in the E-step is of $O(dkN)$ complexity. This efficiency is due to the fact that all the latent variables share the same covariance matrix, which only needs to be computed once. The computational complexity of the M-step is dominated by the inversion of a $dk\times dk$ matrix in the computation of the estimate for $\mathbf{B}$. This is normally of $O(d^3k^3)$ complexity, but since the matrix in question is sparse substantial computationally savings can be achieved by an efficient numerical implementation.

As is the case for PPCA, the resulting constrained principle vectors are not orthogonal. They do, however, span the subspace of interest. If required, the principle vectors can be orthogonalised in a post-processing step. Such an orthogonality constraint can also be built into the model in the form of an additional component in the prior distribution.

4. Experiments and Results

This section illustrates the constrained subspace modelling approach and compares it with unconstrained PCA on two problems, one from speech analysis (Section 4.1) and one from computer vision (Section 4.2). The results for unconstrained PCA were obtained by applying the same algorithm, but setting the prior to be vague, as discussed in Section 2.2. Recall that in this case PPCA is recovered.

4.1. Spectral Speech Model

An important problem in speech analysis is speech compression, *i.e.* finding a lower bit rate representation for the raw speech samples. Compression aims to reduce the bit rate as much as possible without affecting the intelligibility of the decompressed speech. Most compression schemes currently available are model based. Such schemes perform compression by estimating the parameters and residuals for some generative parametric speech model, the autoregressive (AR) process [10] being a popular choice. The model parameters and residuals then constitute the compressed speech representation. The original speech is recovered by running the model in generative mode with the estimated model parameters and residuals.

As opposed to a model based strategy the objective here is to devise an example based compression scheme, where the compressed speech representation is learned from actual human speech samples. More specifically, using a large number of speech spectra the aim is to learn a lower dimensional subspace model that captures the salient information inherent in speech signals. For this purpose a short time processing strategy is adopted here. Such a strategy processes speech signals in (possibly overlapping) blocks of d samples, with d chosen such that the signal is approximately stationary within a block. For each block a spectral representation of the speech is obtained by computing the

Discrete Cosine Transform (DCT) of the speech samples. For a block of speech samples $\mathbf{s} \in \mathbb{R}^d$ the corresponding DCT $\mathbf{x} \in \mathbb{R}^d$ can be expressed in vector-matrix form as $\mathbf{x} = \mathbf{W}\mathbf{s}$, where $\mathbf{W} \in \mathbb{R}^{d \times d}$ is the transform matrix with elements $w_{ij} = c_i \cos(0.5\pi(2j+1)i/d)$, $i, j = 0 \cdots d-1$, with $c_0 = \sqrt{1/d}$ and $c_i = \sqrt{2/d}$, $i = 1 \cdots d-1$. Since $\mathbf{W}$ is an orthogonal matrix the inverse DCT is straightforwardly obtained as $\mathbf{s} = \mathbf{W}^{\mathrm{T}}\mathbf{x}$.

As it stands the DCT representation achieves no data reduction, retaining the same number of coefficients as the number of speech samples. However, intuition suggests that human speech is confined to only a small region within the entire d-dimensional DCT space, and it is the objective here to find the linear subspace that best preserves the perceptual qualities of human speech. Once the subspace parameters $(\mathbf{B}, \boldsymbol{\mu}, \sigma^2)$ are computed compression and decompression of the speech blocks can be performed according to

$$\textbf{Compress}: \quad \widehat{\boldsymbol{\alpha}} = (\mathbf{B}^{\mathrm{T}}\mathbf{B} + \sigma^2\mathbf{I}_k)^{-1}\mathbf{B}^{\mathrm{T}}(\mathbf{W}\mathbf{s} - \boldsymbol{\mu})$$

$$\textbf{Decompress}: \quad \widehat{\mathbf{s}} = \mathbf{W}^{\mathrm{T}}(\mathbf{B}\widehat{\boldsymbol{\alpha}} + \mu).$$

These equations are the MAP solutions for the subspace coefficients and reconstructed data, respectively, and minimise the data reconstruction error, defined as $e = (\mathbf{s} - \widehat{\mathbf{s}})^{\mathrm{T}}(\mathbf{s} - \widehat{\mathbf{s}})$. When processing longer utterances the reconstructed speech blocks are combined using the overlap-add method.

For the data $N = 10,000$ DCT spectra were uniformly randomly extracted from 20 utterances, 2 by each of 10 speakers, 5 being male and 5 being female, taken from the TIMIT database [9]. Utterances in this database are by speakers from across the United States, and are digitised at a sampling rate of 16 kHz and a precision of 16 bits per sample. The block size was set to $d = 128$ samples, corresponding to an analysis window of 8 ms. The utterances selected for this experiment are such that they capture the approximate relative proportions of phonemes in standard United States English.

Figure 1 (left) shows the resulting subspace of size $k = 30$ when applying unconstrained PCA to the data. It captures 86% of the variance present in the data. On closer examination of the principal components it is evident that these focus almost exclusively on the variation in the low frequencies, and capture very little of the variation in the mid frequencies, and virtually none of the variation in the high frequencies. This is due to the fact that most of the signal energy is carried in the low frequencies. The poor modelling of the signal in the mid and high frequencies is further exemplified by the reconstructed test utterance spectogram in Figure 2 (middle). The corresponding audio signal sounds muffled, as if filtered by a lowpass filter.

Given the perceptual importance of preserving the information in the mid and high frequencies, it is desirable to construct a constrained subspace model to capture the variation in these frequency ranges. To this end the constraining prior in (5) was configured to have twenty unconstrained components to capture the global variation, and five components each in the mid and high frequency ranges. The dimensionality of the resulting subspace is thus the same as in the unconstrained case. For a particular principal component the constraining effect was achieved by setting the prior mean to zero, and associating relatively large variances (10^{-3}) with the target frequency ranges, and small variances (10^{-6}) with frequencies outside the target ranges.

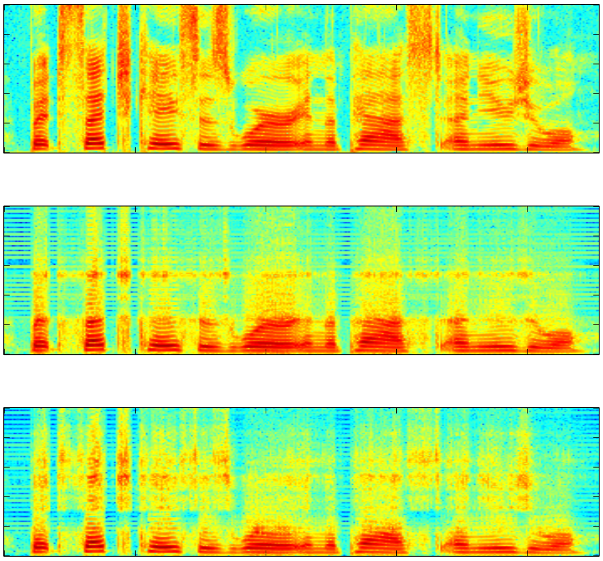

Figure 2. Test utterance spectograms. (Top) Spectogram for the test utterance “But such cases were in the past unusual.” by a female American speaker. (Middle) Spectogram for the unconstrained reconstruction. The loss of high and mid frequency information is evident. The audio signal sounds muffled, as if filtered by a lowpass filter. (Bottom) Spectogram for the constrained reconstruction. The constrained subspace is much more successful at preserving the mid and high frequency information. The audio signal is also of a higher perceptual quality, retaining the crispness of the fricative and plosive sounds.

The resulting subspace is depicted in Figure 1 (right). It captures 82% of the variance present in the data, which is only slightly less than in the unconstrained case. More importantly, however, is the fact that the subspace successfully captures the significant variation in the mid and high frequencies. The spectogram of the reconstructed test utterance using this subspace is shown in Figure 2 (bottom). Comparing this with the unconstrained reconstruction it is

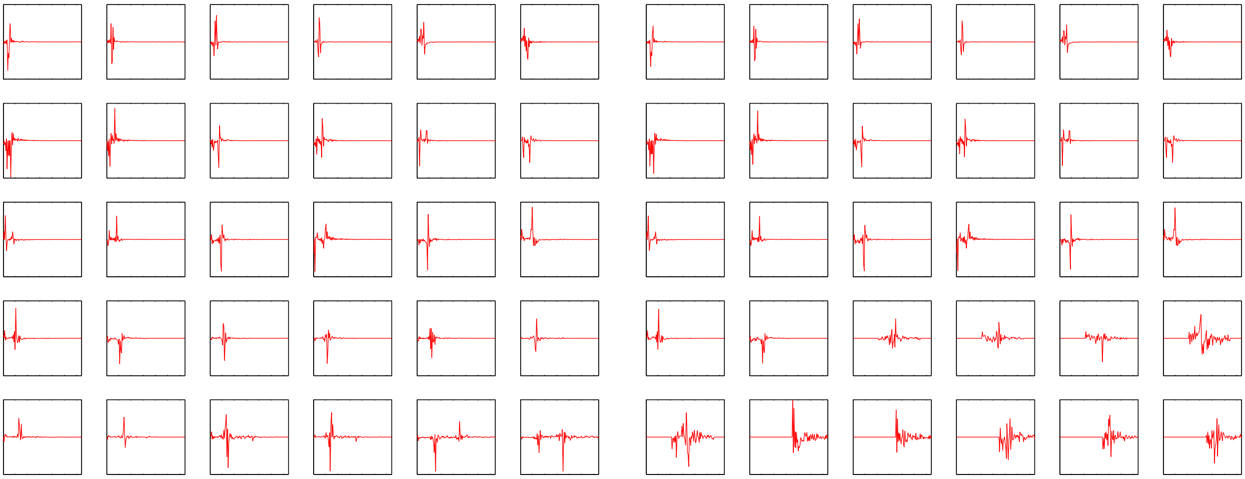

Figure 1. Spectral speech subspaces. (Left) Unconstrained subspace. The unconstrained subspace focuses almost exclusively on the low frequencies. Only the final two principal components capture some of the variation in the mid frequencies. (Right) Constrained subspace. The first 20 principal components are unconstrained, and correspond roughly to the first 20 principal components of the unconstrained subspace. Of the final 10 principal components 5 are constrained to focus on the mid frequencies, and the remaining 5, on the high frequencies.

evident that the constrained modelling approach is much more successful at preserving the mid and high frequency information. The corresponding audio signal is of a higher perceptual quality, retaining the crispness of the fricative and plosive sounds.

4.2. Face Model

Using linear PCA to model the shape and appearance of faces is common in computer vision applications (see *e.g.* [3, 6]). The standard approach has been extended in several ways to deal with the difficulties associated with face modelling. One example is Robust PCA [8], which effectively handles both intra and inter sample outliers.

One desirable feature when using PCA to model faces is to isolate the deformations in certain target regions, such as the eyes and the mouth. These regions can then be reconstructed or animated separately, without affecting the other regions, leading to a more compact and meaningful face model. Previous attempts to achieve this effect include Sparse PCA [5], which encourages the principal components to be sparse (*i.e.* to have few non-zero elements), thus focusing on local regions. However, no constraint is placed on the location of the regions. In [4] the face images are segmented by hand, and an independent subspace is constructed for each region individually. Also important here is the local feature analysis framework [12], that constructs local feature kernels by post-processing the output of standard PCA.

Isolation of key regions in face images can be elegantly achieved using the constrained modelling approach introduced in Section 2. To illustrate this the approach is applied to capture the deformations in certain key regions, namely the eyes and the mouth. For the data the AT&T face database is used, with the individual images subsampled to have a common size of 28 by 23. The database consists of 400 face images, 10 for each of 40 subjects, acquired under roughly similar viewing conditions. Apart from the differences between subjects, the variation in the data is mostly due to variations in expression, and slight variations in pose.

Figure 3 (top) shows the principal components for a $k = 29$ dimensional subspace when applying unconstrained PCA to the data. This subspace captures 84% of the variance present in the data. The first few principal components are relatively smooth, and represent global correlations. As the variance decreases the principal components represent finer spatial structure, but remain global in space.

Figure 3 (bottom) shows the subspace for the constrained modelling approach, where 5 principal components have been constrained to focus on the mouth region, and 7 principal components, on the region around the eyes. The constraining effect has been achieved in the same way as for the speech subspace in Section 4.1. The resulting subspace captures 81% of the variance present in the data, which is only slightly less than the unconstrained model, but leads to a more meaningful face representation. Figure 4 shows

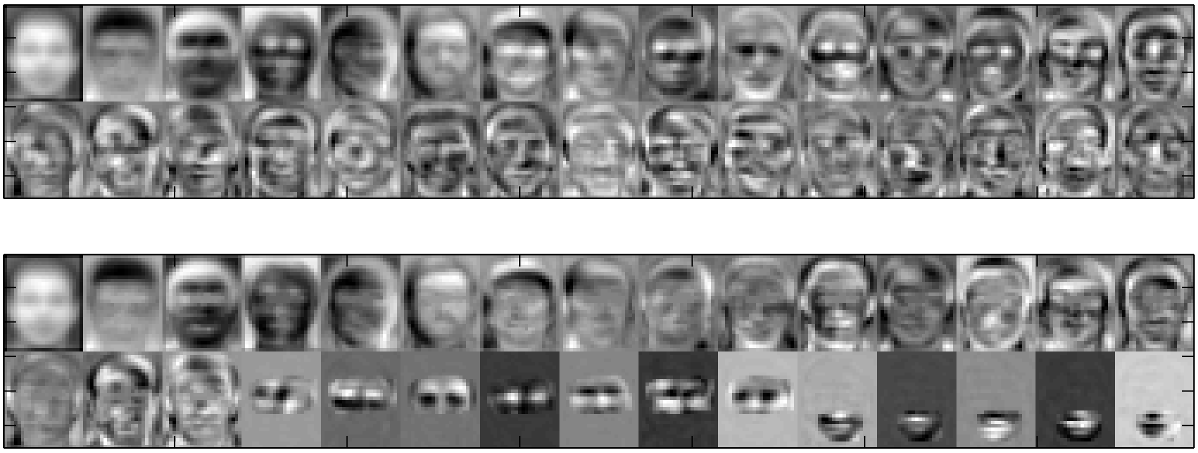

Figure 3. Face subspaces. (Top) Unconstrained subspace. The first few principal components are smooth, and represent global correlations. The remaining principal components represent finer spatial structure, but remain global in space. (Bottom) Constrained subspace, with 5 principal components constrained to focus on the mouth region, and 7 principal components, on the region around the eyes. The unconstrained principal components correspond roughly to the ones for the unconstrained subspace, but without the same degree of variation in the eye and mouth regions.

an example where the change of expression in a novel face (not in the training set) is captured predominantly by the components constrained to the mouth and eye regions.

5. Conclusions

This paper introduced a constrained subspace modelling strategy by augmenting the PPCA likelihood with a prior to achieve the constraining effect. An efficient EM algorithm was presented to estimate the parameters of the resulting subspace model. The constrained modelling approach is especially useful in problems where the data is approximately isotropic along the lesser principal components. These can then be constrained to be more meaningful in the context of the problem being investigated without any significant loss in the modelling fidelity. This effect was evident in the problems investigated here. In both cases the constrained subspaces were more interpretable, with only a slight reduction in the proportion of variance captured. In the speech analysis application the reconstructed signals were of a higher perceptual quality. In the face modelling setting the constrained approach allowed the deformations in the eye and mouth regions to be compactly represented and isolated from the remaining global deformations.

It may be argued that the results obtained by the constrained subspace modelling approach may also be achieved by simpler approaches, such as applying standard PCA independently to subregions of the space, or by rescaling the data prior to the application of standard PCA. The former may indeed by viewed as a special case of the constrained approach, where the regions are non-overlapping and span the space. The proposed framework is more general, and can accommodate arbitrary regions that interact in arbitrary ways. The same holds true for the rescaling of the data. Apart from being less elegant, in that it modifies the data, it can only be applied to mutually exclusive regions in the space.

References

[1] C. M. Bishop. Bayesian PCA. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11, pages 382–388. MIT Press, 1999.

[2] C. M. Bishop and J. M. Winn. Non-linear Bayesian image modelling. In *Proc. Europ. Conf. Computer Vision*, pages I: 3–17, 2000.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proc. ACM Siggraph*, pages 187–194, 1999.

[4] D. Capel and A. Zisserman. Super-resolution from multiple views using learnt image models. In *Proc. Conf. Comp. Vision Pattern Rec.*, 2001.

[5] C. Chennubhotla and A. Jepson. Sparse PCA: Extracting multi-scale structure from data. In *Proc. Int. Conf. Computer Vision*, pages I: 641–647, 2001.

[6] T. F. Cootes, K. Cooper, C. J. Taylor, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

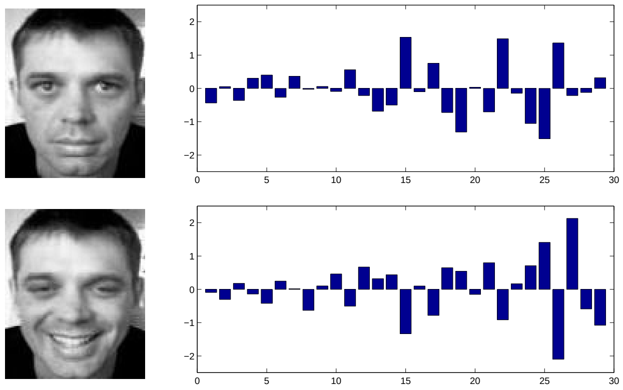

Figure 4. Face compression. (Top) Reference face and its mapping onto the constrained subspace. (Bottom) Novel expression and the resulting change in the subspace coefficients, compared to the reference face. The coefficients in the mouth region (25 to 29) capture 41% of the change, and those in the region of the eyes (18 to 24), 22%.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[8] F. D. la Torre and M. J. Black. Robust principal component analysis for computer vision. In *Proc. Int. Conf. Computer Vision*, pages I: 362–369, 2001.

[9] L. F. Lamel, R. H. Kassel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 100–109, 1986.

[10] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[11] R. M. Neal. Assessing relevance determination methods using DELVE. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, pages 97–129. Springer-Verlag, 1998.

[12] P. S. Penev and J. J. Atick. Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.

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

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