Smart PCA

Yi Zhang
Machine Learning Department
Carnegie Mellon University
yizhang1@cs.cmu.edu

Abstract

PCA can be smarter and makes more sensible projections. In this paper, we propose smart PCA, an extension to standard PCA to regularize and incorporate external knowledge into model estimation. Based on the probabilistic interpretation of PCA, the inverse Wishart distribution can be used as the informative conjugate prior for the population covariance, and useful knowledge is carried by the prior hyperparameters. We design the hyperparameters to smoothly combine the information from both the domain knowledge and the data itself. The Bayesian point estimation of principal components is in closed form. In empirical studies, smart PCA shows clear improvement on three different criteria: image reconstruction errors, the perceptual quality of the reconstructed images, and the pattern recognition performance.

1 Introduction

Principal components analysis (PCA) [Jolliffe, 2002] is a standard technique for dimensionality reduction and feature extraction. It is widely used in multivariate statistics, machine learning, image processing and computer vision, where hundreds or even thousands of features are confronted. Given infinite data, PCA discovers optimal principal components in terms of variance maximization, or equivalently, reconstruction error minimization [Jolliffe, 2002]. However, in situations with large numbers of features and only a limited amount of observations, covariance estimation is difficult and PCA will inevitably overfit the sample covariance. Another way to understand this problem is the probabilistic interpretation of PCA [Tipping and Bishop, 1999; Roweis, 1997], where principal components corresponds to the maximum likelihood estimation (MLE) of parameters in a latent variable model, and thus tend to overfit limited observations in high-dimensional space.

To tackle this problem, we propose to regularize and incorporate external knowledge into PCA via the inverse Wishart prior, and as a result, PCA can be smarter, make more sensible projections, and construct more useful features. The basis of our work includes the probabilistic interpretation of PCA [Tipping and Bishop, 1999; Roweis, 1997], which views PCA as a specific case of factor analysis with isotropic Gaussian noise, and the use of the inverse Wishart distribution as the natural conjugate prior for the covariance matrix in multivariate normal distribution [Gelman *et al.*, 2003], which has been recently investigated by researchers in statistics [Brown *et al.*, 2000; Press, 2005], machine learning [Klami and Kaski, 2007], image processing and computer vision [Smidl *et al.*, 2001; Wood *et al.*, 2006]. Based on previous work, a natural way to improve PCA is to incorporate external knowledge through the prior distribution on model parameters, which is the concern of this paper. External knowledge can be embedded into PCA through the inverse Wishart distribution, and the result of using such a conjugate prior is a straightforward Bayesian point estimation of principal components. Given external knowledge in terms of feature relevance or distance, we design the prior hyperparameters to include the information from both external knowledge and the data itself, so that the resulting prior is informative and robust. We discuss the choice of feature distance for image processing. We hope this will encourage domain experts to design feature distance or relevance functions suitable for their own domains. Empirical studies of smart PCA show promising results on overfitting control and feature construction, in terms of image reconstruction errors, the perceptual quality of reconstructed images, and pattern recognition performance.

2 Smart PCA

With large numbers of features and only limited amount of data, PCA will overfit the sample covariance as an estimation of the population covariance. To address this problem, a natural way is to regularize and incorporate external knowledge through a prior based on the probabilistic interpretation of PCA [Tipping and Bishop, 1999; Roweis, 1997]. In this section, we introduce smart PCA. Section 2.1 reviews probabilistic PCA. Section 2.2 discusses the use of inverse Wishart distribution as the conjugate prior of probabilistic PCA and the Bayesian point estimation of principal components. Section 2.3 focuses on using external knowledge in forms of feature distance to construct the hyperparameters of the prior.

2.1 Probabilistic PCA

The probabilistic interpretation of PCA [Tipping and Bishop, 1999; Roweis, 1997] considers PCA as a special case of factor analysis with isotropic Gaussian noise. In this section

we briefly review this probabilistic framework. It is assumed that each observed p-dimensional vector $\mathbf{y}$ was transformed from k-dimensional latent variables $\mathbf{x}$ where $k \leq p$. This transformation is determined by a $p \times k$ matrix $\mathbf{W}$ and a p-dimensional isotropic Gaussian noise ϵ:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \epsilon \tag{1}$$

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{3}$$

Note that latent variables $\mathbf{x}$ follow standard k-d normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and isotropic noise ϵ is p-d normal distribution $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. This formula also omits the population mean μ[1]. Since $\mathbf{x}$ and ϵ are independent Gaussian, observed variables $\mathbf{y}$ also obeys multivariate normal distribution:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^\mathbf{T} + \sigma^2\mathbf{I}) \tag{4}$$

To obtain standard PCA, we take the limit $\sigma^2 \to 0^+$:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{5}$$
$$\mathbf{\Sigma} = \lim_{\sigma^2 \to 0^+} \mathbf{W}\mathbf{W}^\mathbf{T} + \sigma^2\mathbf{I}$$

where $\mathbf{\Sigma} = \lim_{\sigma^2 \to 0^+} \mathbf{W}\mathbf{W}^\mathbf{T} + \sigma^2\mathbf{I}$ is the population covariance. By limiting $\sigma^2 \to 0^+$, we rely on $\mathbf{W}$ to explain the observed variables $\mathbf{y}$ (i.e., $\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^\mathbf{T}$). Given N observations $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, the log-likelihood is:

$$\mathcal{L}(\mathbf{\Sigma}|\mathbf{Y}) \propto -\frac{N}{2}\{\ln|\mathbf{\Sigma}| + tr(\mathbf{\Sigma}^{-1}\mathbf{S})\} \tag{6}$$

where $\mathbf{S}$ is the *sample* covariance matrix:

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i\mathbf{y}_i^T \tag{7}$$

This log-likelihood is defined on $\mathbf{\Sigma} = \lim_{\sigma^2 \to 0^+} \mathbf{W}\mathbf{W}^\mathbf{T} + \sigma^2\mathbf{I}$ and is maximized when [Tipping and Bishop, 1999]:

$$\mathbf{W}_{ML} = \lim_{\sigma^2 \to 0^+} \mathbf{U}_k(\mathbf{\Lambda}_k - \sigma^2\mathbf{I})^{\frac{1}{2}} \tag{8}$$

where the columns of $p \times k$ matrix $\mathbf{U}_k$ are k eigenvectors (i.e., principal components) of the sample covariance $\mathbf{S}$ which correspond to the k largest eigenvalues, and $\mathbf{\Lambda}_k$ is the $k \times k$ diagonal matrix containing these k eigenvalues. In this sense, principal components of standard PCA are recovered, and the corresponding projection and reconstruction get a reasonable probabilistic interpretation [Tipping and Bishop, 1999; Roweis, 1997]. More specifically, given an observed p-d vector $\mathbf{y}$, its orthogonal projection into the k-d latent space can be explained as the conditional expectation of $\mathbf{x}$ given $\mathbf{y}$:

$$E(\mathbf{x}|\mathbf{y}) = \lim_{\sigma^2 \to 0^+} (\mathbf{W}^\mathbf{T}\mathbf{W} + \sigma^2\mathbf{I})^{-1}\mathbf{W}^\mathbf{T}\mathbf{y} \tag{9}$$

Also, the reconstruction from a k-d latent variable $\mathbf{x}$ to the p-d space corresponds to the conditional expectation of $\mathbf{y}$:

$$E(\mathbf{y}|\mathbf{x}) = \mathbf{W}\mathbf{x} \tag{10}$$

[1]The factor analysis usually includes the population mean μ: $\mathbf{y} = \mathbf{W}\mathbf{x} + \mu + \epsilon$. The optimal estimation for μ is the sample mean. In practice, people firstly subtract the sample mean from the observations and then assume zero mean [Roweis, 1997].

2.2 Inverse-Wishart distribution and smart PCA

In this section we discuss the inverse Wishart distribution [Gelman *et al.*, 2003] as the prior for the population covariance in probabilistic PCA and the resulting Bayesian point estimation of principal components.

Consider the $p \times p$ population covariance $\mathbf{\Sigma}$ in (5). We say that $\mathbf{\Sigma}$ follows an inverse Wishart distribution $IW_p(\mathbf{G}, \nu)$ with positive definite scale matrix $\mathbf{G}$ and degree of freedom $\nu > 2p$ if [Press, 2005; Gelman *et al.*, 2003]:

$$p(\mathbf{\Sigma}|\mathbf{G}, \nu) = \frac{c_0|\mathbf{G}|^{(\nu-p-1)/2}}{|\mathbf{\Sigma}|^{\nu/2}} exp(-\frac{1}{2}tr(\mathbf{\Sigma}^{-1}\mathbf{G})) \tag{11}$$

where c_0 is a normalization constant.

The mode of an inverse Wishart distribution is given by:

$$Mode(\mathbf{\Sigma}|\mathbf{G}, \nu) = \mathbf{G}/\nu \tag{12}$$

Adopted from [Chen, 1979], we reparametrize the inverse Wishart distribution as:

$$\mathbf{\Sigma}|\mathbf{\Omega}, \nu \sim IW_p(\nu\mathbf{\Omega}, \nu) \tag{13}$$

where ν is also the degree of freedom, and $\mathbf{\Omega}$ represents the *prespecified structural information* about $\mathbf{\Sigma}$ in that:

$$Mode(\mathbf{\Sigma}|\mathbf{\Omega}, \nu) = \nu\mathbf{\Omega}/\nu = \mathbf{\Omega} \tag{14}$$

Inverse Wishart distribution is the conjugate prior of the population covariance in multivariate normal distribution [Press, 2005; Gelman *et al.*, 2003]. Suppose N observations $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ with $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is assumed to follow an inverse Wishart prior $IW_p(\nu\mathbf{\Omega}, \nu)$. The posterior distribution of $\mathbf{\Sigma}$ still follows an inverse Wishart:

$$\mathbf{\Sigma}|\mathbf{\Omega}, \nu, \mathbf{Y} \sim IW_p(\nu^\star\mathbf{\Omega}^\star, \nu^\star) \tag{15}$$

where the parameters of the posterior distribution are:

$$\nu^\star = \nu + N \tag{16}$$

$$\mathbf{\Omega}^\star = (\frac{N}{N+\nu})\mathbf{S} + (\frac{\nu}{N+\nu})\mathbf{\Omega} \tag{17}$$

Note that $\mathbf{\Omega}^\star$ is a weighted combination of the prior hyperparameter $\mathbf{\Omega}$ and the sample covariance $\mathbf{S}$, and is also the mode of the posterior (inverse Wishart) distribution.

Recall in (8) that the solution to standard PCA is found by maximizing the log-likelihood (6) with respect to $\mathbf{W}$. Given an informative prior $IW_p(\nu\mathbf{\Omega}, \nu)$ on $\mathbf{\Sigma}$, it is natural that maximum likelihood estimation (MLE) is replaced by the maximum a posteriori (MAP) estimation. In other words, the Bayesian point estimation of principal components corresponds to maximizing the posterior density of $\mathbf{\Sigma}$ in (15) with respect to $\mathbf{W}$ ($\mathbf{\Sigma} = \lim_{\sigma^2 \to 0^+} \mathbf{W}\mathbf{W}^\mathbf{T} + \sigma^2\mathbf{I}$). Based on (11), the log-posterior-density of (15) can be written as:

$$\mathcal{LP}(\mathbf{\Sigma}|\mathbf{\Omega}^\star, \nu^\star) \propto -\frac{\nu^\star}{2}\{\ln|\mathbf{\Sigma}| + tr(\mathbf{\Sigma}^{-1}\mathbf{\Omega}^\star)\} \tag{18}$$

Note that this log-posterior-density is exactly the same form as the log-likelihood in (6), except that $\mathbf{S}$ and N are replaced by $\mathbf{\Omega}^\star$ and $\nu^\star$. In this sense, maximum a posteriori (MAP) estimation of the $p \times k$ matrix $\mathbf{W}$ is given by:

$$\mathbf{W}_{MAP} = \lim_{\sigma^2 \to 0^+} \mathbf{U}_k^\star(\mathbf{\Lambda}_k^\star - \sigma^2\mathbf{I})^{\frac{1}{2}} \tag{19}$$

This is the same form as (8), but here $\mathbf{U}_k^\star$ and $\mathbf{\Lambda}_k^\star$ are the first k principal eigenvectors and eigenvalues of $\mathbf{\Omega}^\star$ in (17) rather than sample covariance $\mathbf{S}$. To this end, the k principal components are discovered in a Bayesian setting and this process is the basic framework of ***Smart PCA***. Also, as in (9) and (10), the projection and reconstruction under smart PCA still have reasonable probabilistic interpretations.

2.3 Hyperparameter design

This section focuses on the hyperparameters of the inverse Wishart prior in (13). In smart PCA, principal components are extracted from $\mathbf{\Omega}^\star$, which is defined in (17). It is a weighted combination of the sample covariance $\mathbf{S}$ and the hyperparameter $\mathbf{\Omega}$, where $\mathbf{\Omega}$ represents the prespecified covariance structure. The strength of this combination is controlled by ν, the degree of freedom of the prior. For the density of (13) to be valid, ν need to be an integer larger than $2p$. But in practice, there is no reason for follow this restriction. In smart PCA, ν represents how much we trust the hyperparameter $\mathbf{\Omega}$, and it can be set empirically as any non-negative real number. When it is set as zero, $\mathbf{\Omega}^\star$ is exactly the sample covariance $\mathbf{S}$, and smart PCA is equal to standard PCA.

A central problem of smart PCA is the construction of hyperparameter $\mathbf{\Omega}$, which represents prespecific structural information about covariance. Domain knowledge can be incorporated into $\mathbf{\Omega}$ so that smart PCA can extract more sensible principal components. In this section, we assume the general form of the external knowledge is some feature distance (or relevance) function:

$$d(\mathbf{f_i}, \mathbf{f_j}) \tag{20}$$

where $\mathbf{f_i}$ and $\mathbf{f_j}$ are the ith and jth features (i.e., dimensions) in the observation space. The design of domain-dependent feature distance will be mentioned in Section 3.

It is challenging to find a direct mapping from a given feature distance function $d()$ to the covariance structure $\mathbf{\Omega}$. The feature distance can be in any scale, and covariance is also a scale-free measure. For example, simply scaling up the value of each feature will increase the covariance, but the feature distance based on domain knowledge will not change since the features themselves don't change. To overcome this problem, we combine both knowledge from the feature distance function and information from the data itself to construct $\mathbf{\Omega}$.

Firstly, a covariance matrix can be decomposed into standard deviation and correlation [Gelman *et al.*, 2003]. Inspired by this, the $p \times p$ covariance structure $\mathbf{\Omega}$ can be written as:

$$\mathbf{\Omega} = \mathbf{VCV} \tag{21}$$

where $\mathbf{V} = Diag(\sigma_1, \sigma_2, \ldots, \sigma_p)$ is a $p \times p$ diagonal matrix, and $\mathbf{C}$ is a $p \times p$ positive definite matrix since $\mathbf{\Omega}$ is required to be definite positive as in inverse Wishart distribution. The intuition of this decomposition is: the diagonal elements of $\mathbf{V}$ contain information about standard deviation and $\mathbf{C}$ represents the prespecified *correlation* structure. In this sense, the construction of $\mathbf{\Omega}$ consists of two parts: $\mathbf{V}$ and $\mathbf{C}$.

Secondly, $\mathbf{V} = Diag(\sigma_1, \sigma_2, \ldots, \sigma_p)$ is estimated from data, i.e., it is set as the maximum likelihood estimation of the standard deviation in each dimension. Although the MLE estimation of the population covariance in p-dimensional feature space involves $O(p^2)$ parameters and tends to overfit the limited observations, the estimation of standard deviations involves only p parameters and is much more stable.

The third step is to estimate $\mathbf{C}$, the prespecified correlation structure. Adapted from [Yaglom, 1987], a general form of the correlation function given the feature distance d is:

$$C_{ij} = exp(-\frac{d(\mathbf{f_i}, \mathbf{f_j})}{\alpha}) \tag{22}$$

where $\alpha > 0$ is a parameter. Similar forms have been used for covariance estimation [Brown *et al.*, 2000; Krupka and Tishby, 2007], but this class of functions are more suitable for estimating correlation since their outputs are in the range $[0, 1]$. Note that α needs to be appropriately specified according to the scale of function d. For example, if α is several orders of magnitude larger than the maximum value of feature distance d, all the elements in $\mathbf{C}$ will be closed to 1, without regard to the information contained in d. To specify α, we compute the following two statistics: 1) ρ_{median}: the median element in the sample *correlation* matrix, which can be readily obtained from the sample covariance matrix $\mathbf{S}$. 2) d_{median}: the median element of the feature distance matrix, which is computed using distance function d. Based on these two statistics, α is set to meet the following criteria:

$$\rho_{median} = exp(-\frac{d_{median}}{\alpha}) \tag{23}$$

Note that ρ_{median} is a robust statistic on sample, characterizing the typical level of observed feature correlation. Also, d_{median} is a robust statistic on external knowledge, describing the typical level of feature distance. In this sense, the α connecting these two statistics is a reliable choice.

3 Feature Distance Functions for Image Processing

Section 2 proposes the general framework of smart PCA. The last problem left is to design domain-dependent feature distances with external knowledge. This section provides two examples for image precessing and computer vision.

3.1 Spatial distance of pixels

Krupka *et al.* [Krupka and Tishby, 2007] propose a general function to measure feature distance. Given a vector of meta-features describing each feature, feature distance can be a distance function on meta-features. This is suitable for computer vision tasks where features are pixels and meta-features are the location of pixels. In our experiments, we use the Euclidian distance of the location of pixels as the feature distance.

3.2 Geodesic distance of pixels

Intuitively, pixels in the same stroke are highly relevant and likely to be more correlated. On the other hand, pixels depicting two distinct entities in an image tend to be independent. Inspired by Isomap [Tenenbaum *et al.*, 2000], we propose a feature distance to capture this information. The proposed distance measure makes use of both meta-features (location of pixels) and value of features (e.g., grayscale value), and is computed as the following steps. *1) Define neighborhood*:

the neighborhood of a pixel is defined as the immediate neigh bors in eight possible directions. For pixels not in the bound ary of the image, the eight surrounding pixels are neighbors For pixels in the boundary, only three or five neighbors ex ist. *2) Compute local metric*: given an image, the distanc between two neighboring pixels is computed as the differ ence of their grayscale values. Given a collection of images the local distance of two neighboring pixels is averaged ove. each image. *3) Construct neighborhood graph*: construct a graph where each node denotes a pixel. Connect neighboring pixels by undirected edges and set the weight of each edge as the local distance of two pixels. *4) Compute shortest path distance*: shortest path distance is computed by efficient algorithms [Tenenbaum *et al.*, 2000]. *5) Regularization*: a smal quantity is added to the diagonal elements to make the dis tance matrix definite positive if necessary. Finally, the dis tance of two features (pixels) is their shortest path distance.

4 Empirical Study

Based on a benchmark data set in computer vision, we stud the performance of smart PCA as a tool for dimensionalit reduction and feature construction. We evaluate the performance using three criteria: 1) image reconstruction errors; 2) the perceptual quality of the reconstructed images; 3) pattern recognition (i.e., classification) performance using principal components. With moderate external knowledge as mentioned in section 3, smart PCA shows clear improvement over standard PCA on all these criteria.

4.1 Experimental Settings

The benchmark dataset used in this paper is the Yale face database, which includes 165 image for 15 people and thus 11 images for each person. Images are resized to 32×32 pixels, indicating a feature space of 1024 dimensions. The result of each experiment is average over 10 random runs. In each run, training and testing examples are selected as follows: randomly selecting 6 images for each individual as training data and the rest as testing data, which leads to totally 90 training and 75 testing examples in each random run. In each run, both the population mean and principal components are estimated from training examples and applied to each previously unseen testing example.

The two parameters of smart PCA are $\mathbf{\Omega}$ and ν, which are the hyperparameters for the inverse Wishart prior as shown in (13). The $\mathbf{\Omega}$ is automatically determined as stated in Section 2.3, and ν is empirically determined. Recall in (17) that ν control the strength of covariance prior $\mathbf{\Omega}$ when it is combined with the sample covariance $\mathbf{S}$. To gain a comprehensive understanding of the effect of ν, we test different choices. Given the number of observations (i.e, training examples) N, we test different ν by choosing different $\frac{\nu}{N}$ (i.e., "relative prior strength" w.r.t. the number of training examples) from the following range: $\{0, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0, 1.5, 2.0, 3.0, 5.0, 6.0, 7.0, 8.0, 10.0, 12.0, 15.0\}$. The results will indicate the impact of ν on smart PCA. For example, when ν is set as $1.0N$ (i.e., $\frac{\nu}{N} = 1.0$), the posterior hyperparameter $\mathbf{\Omega}^\star$ in (17) is exactly the average of sample covariance $\mathbf{S}$ and the prior $\mathbf{\Omega}$. Principal components are extracted based on the resulting $\mathbf{\Omega}^\star$, as shown in (19).

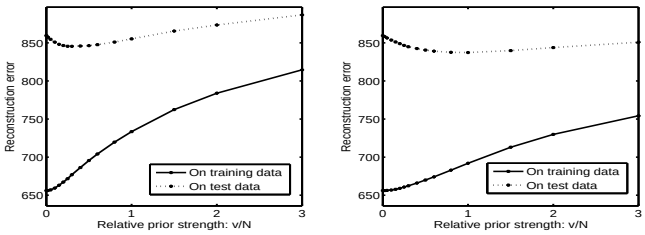

Figure 1: Reconstruction error using 20 principal components. Left: spatial distance used in Smart PCA; right: geodesic distance used in Smart PCA.

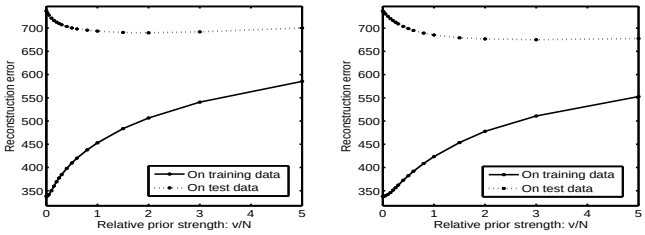

Figure 2: Reconstruction error using 50 principal components. Left: spatial distance used in Smart PCA; right: geodesic distance used in Smart PCA.

We have discussed two distance functions on pixels in Section 3. In our experiments, we study both of them: 1) the spatial (i.e., Euclidian) distance of pixel location (x,y) in the image; 2) the geodesic distance of pixels.

4.2 Empirical Results

In this section, we discuss empirical results based on three criteria: 1) image reconstruction errors; 2) the perceptual quality of the reconstructed images; 3) pattern recognition rates.

Reconstruction errors

It is well known that the principal components found by standard PCA is optimal in terms of reconstruction error minimization. However, this optimality is true only on the training examples, i.e., on the data where the PCA is learned. In our empirical study, we construct PCA and smart PCA from a set of training images, and measure the reconstruction errors on both the training images and a set of unseen testing images. Experimental results show that standard PCA overfits the training examples, and smart PCA corrects this bias. We run experiments with different parameter ν and distance functions, and results are averaged over 10 random runs.

The reconstruction error on an image is defined as the root mean square error (RMSE) over the reconstruction error of all pixels, and the reconstruction error on an set of images is the average over all images in the set. The results on reconstruction errors are shown in Figure 1 and Figure 2, corresponding to using the first 20 and 50 principal components, respectively.

In the figures, the standard PCA corresponds to the choice $\frac{\nu}{N} = 0$. The reconstruction error is measured on both training images and unseen testing images, and shown as the solid

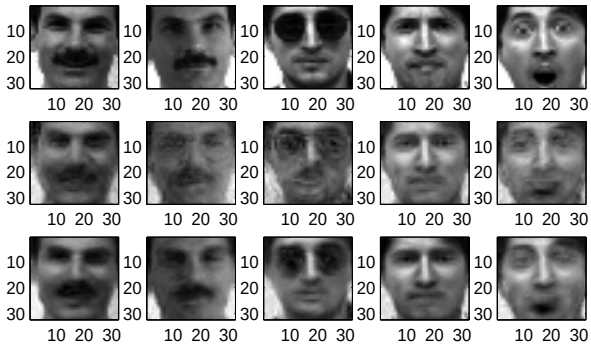

Figure 3: Original images vs. PCA reconstructions using 150 PCs vs. Smart PCA reconstructions using 150 PCs.

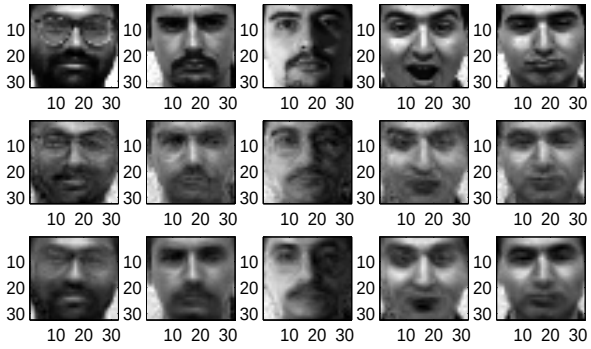

Figure 4: Original images vs. PCA reconstructions using 150 PCs vs. Smart PCA reconstructions using 150 PCs.

curve and dotted curve, respectively. The black dots on each curve is the choices of $\frac{\nu}{N}$ as mentioned in Section 4.1. The curve is interpolated from these dots. We don't display the entire range of $\frac{\nu}{N}$, since certain choices are lack of interest and will change of scale of the plots.

The results of our experiments show that: **1)** Standard PCA does overfit the training data, and smart PCA is able to correct this problem. In our experiments, the best reconstruction error on training images is always given by standard PCA ($\frac{\nu}{N} = 0$), while the lowest reconstruction error on unseen testing images is always obtained by smart PCA ($\frac{\nu}{N} > 0$). **2)** The need for regularization (i.e., Smart PCA) is more evidence when more principal components are estimated. From the figures it is clear that smart PCA provides more improvement over PCA when we reconstruct images using 50 principal components. This is reasonable in that more principal components indicates more coefficients to estimate, which is more likely to suffer from overfitting. **3)** With moderate domain knowledge in the form of feature distance, smart PCA is able to control overfitting and improve reconstruction performance. Well-designed feature distance functions that contain rich domain knowledge lead to better performance of smart PCA. This is supported by our experiments where smart PCA using the geodesic distance of pixels performs better than smart PCA using the spatial distance of pixels.

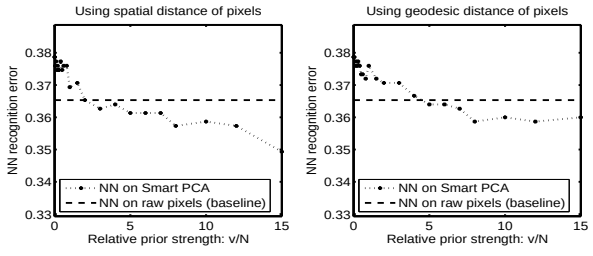

Figure 5: Recognition errors using 50 principal components.

Perceptual quality of reconstructed images

The reconstruction error is not a comprehensive measure for the performance of dimensionality reduction techniques: two reconstructed images with same error can provide very different perceptual qualities. A "smart" model is able to prevent significant perceptual loss even if a considerable amount of information is lost due to dimensionality reduction. Therefore, we also study the perceptual quality of the reconstructed (testing) images using standard PCA and smart PCA. We use 150 principal components to ensure an acceptable perceptual quality for both models (i.e., dimensionality reduced from 1024 to 150). For smart PCA, we report the results using geodesic pixel distance and $\frac{\nu}{N} = 1$. Figure 3 and Figure 4 show a few examples of image reconstruction. Each figure contains 15 images: three images in a column corresponds to a testing example, where the three images from the first row to the third row are the original image, the standard PCA reconstruction, and the smart PCA reconstruction, respectively.

Figure 3 and Figure 4 show that smart PCA leads to reconstructed images with significantly better perceptual quality. As mentioned in Section 3.2, geodesic distance of pixels represents the belief that pixels in the same stroke are relevant and likely to be more correlated in the principal components. As a result, images reconstructed by smart PCA tend to be locally coherent (e.g., mouth, glasses, beard, eyes, etc.) due to regularization on the estimation of principal components.

Pattern recognition errors

Another benefit of dimension reduction is to prevent the curse of dimensionality on statistical modeling. We also include a preliminary study on pattern recognition. The problem is a very hard 15-class recognition task. We train the Nearest-Neighbor (NN) classifier on 50 principal components of training examples, with different $\frac{\nu}{N}$. The experiment is repeated and averaged over 10 random runs. The results are shown in Figure 5. Note that $\frac{\nu}{N} = 0$ corresponds to standard PCA, and the dashed horizonal line is the performance of NN classifier on raw pixels (i.e., baseline without dimension reduction). The results indicate that PCA can construct useful features (i.e., beat the baseline model) only after being regularized.

5 Conclusion

PCA can be smarter and makes more sensible projections. In this paper, we propose to regularize and incorporate external knowledge into PCA. Based on the probabilistic interpretation of PCA, the inverse Wishart distribution is used as the

conjugate prior for the population covariance in factor analysis, and domain knowledge can be transferred by the prior hyperparameters. We design the hyperparameters to combine the information from both the external knowledge and the data itself, so that the prior is informative and robust. The Bayesian point estimation of principal components is in closed form. Empirical studies show clear improvement on image reconstruction errors, the perceptual quality of the reconstructed images, and pattern recognition performance, indicating that Smart PCA is a useful alternative to standard PCA for dimensionality reduction and feature construction.

6 Related work

This paper is mainly built on probabilistic principal component analysis (PPCA), which has been independently proposed by different researchers [Tipping and Bishop, 1999; Roweis, 1997]. Tipping *et al.* focused on the probabilistic interpretation of PCA, while Roweis mainly concentrated on the resulting EM algorithm. In addition, P-PCA has been extended to adaptively select the number of principal components via Bayesian approaches [Bishop, 1998; Minka, 2000]. In this paper, we deal with another problem, that is, how to make PCA smarter and make more sensible projections by incorporating external knowledge.

Another line of research that inspires our work is the use of external knowledge in supervised learning [Hastie *et al.*, 1995; Krupka and Tishby, 2007]. Hastie *et al.* [Hastie *et al.*, 1995] proposed a penalized version of linear discriminant analysis, where the penalty represents the spatial smoothness of model parameters. Krupka *et al.* [Krupka and Tishby, 2007] incorporated the prior information from meta-features into SVM, where the similarity of meta-features indicates the relevance of corresponding features, and the resulting covariance matrix is directly embedded into SVM optimization.

Using the inverse Wishart distribution as the conjugate prior of the population covariance in multivariate normal distribution has a long history in Bayesian statistics [Gelman *et al.*, 2003]. In probabilistic PCA, the observations are explained by a multivariate normal distribution conditional on the population covariance (which is further determined by principal components). In this sense, smart PCA is built on the classic work on Bayesian estimation of covariance matrix [Chen, 1979; Brown *et al.*, 2000; Press, 2005], although the goal of PCA is not to estimate the covariance but the principal components responsible for the covariance.

Recently, the inverse Wishart distribution has been used by researchers in machine learning, imaging processing and computer vision [Klami and Kaski, 2007; Smidl *et al.*, 2001; Wood *et al.*, 2006] as the conjugate prior for multivariate normal distribution in different models. In [Klami and Kaski, 2007], the authors extend probabilistic canonical correlation analysis (PCCA) to enable a full Bayesian treatment via specifying conjugate priors for model parameters of PCCA, where the inverse Wishart distribution is used as the prior for system covariance. In [Wood *et al.*, 2006], the inverse Wishart distribution is used to extend Gaussian Mixture Models to encode prior experience about the shape and position of the mixture components. Researchers also use the inverse Wishart distribution to model the noise covariance in factor analysis for data with a low signal-to-noise ratio, inhomogeneous, or correlated noise [Smidl *et al.*, 2001]. In other words, the noise ϵ in eq. (1) is not assumed to be isotropic or negligible in scale, and thus eq. (3) is no longer valid and the covariance $\sigma^2\mathbf{I}$ is replaced by a general covariance $\mathbf{C}$, which is modeled by an inverse Wishart distribution.

References

[Bishop, 1998] C. M. Bishop. Bayesian PCA. In *NIPS*, pages 382–388, 1998.

[Brown *et al.*, 2000] P. J. Brown, T. Fearn, and M. S. Haque. Discrimination with Many Variables. *J. Am. Statist. Assoc.*, 94(448):1320–1329, 2000.

[Chen, 1979] C. Chen. Bayesian Inference for a Normal Dispersion Matrix and its Application to Stochastic Multiple Regression Analysis. *J. R. Statist. Soc. B*, 41(2):235–248, 1979.

[Gelman *et al.*, 2003] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis, 2nd ed.* Chapman&Hall/CRC, 2003.

[Hastie *et al.*, 1995] T. Hastie, A. Buja, and R. Tibshirani. Penalized Discriminant Analysis. *The Annals of Statistics*, 23(1):73–102, 1995.

[Jolliffe, 2002] I. T. Jolliffe. *Principal Component Analysis, 2nd ed.* Springer, 2002.

[Klami and Kaski, 2007] A. Klami and S. Kaski. Local Dependent Components. In *ICML*, pages 425–432, 2007.

[Krupka and Tishby, 2007] E. Krupka and N. Tishby. Incorporating Prior Knowledge on Features into Learning. In *AISTATS*, pages 227–234, 2007.

[Minka, 2000] T. P. Minka. Automatic choice of dimensionality for PCA. In *NIPS*, pages 598–604, 2000.

[Press, 2005] S. J. Press. *Applied Multivariate Analysis, 2nd ed.* Dover Publications, Inc., 2005.

[Roweis, 1997] S. Roweis. EM Algorithms for PCA and SPCA. In *NIPS*, pages 626–632, 1997.

[Smidl *et al.*, 2001] V. Smidl, M. Karny, and A. Quinn. On Prior Information in Principal Component Analysis. In *Irish Signal and Systems Conference*, 2001.

[Tenenbaum *et al.*, 2000] J. B. Tenenbaum, V. D. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2323, 2000.

[Tipping and Bishop, 1999] M. E. Tipping and C. M. Bishop. Probabilistic Principal Component Analysis. *J. R. Statist. Soc. B*, 61(3):611–622, 1999.

[Wood *et al.*, 2006] F. Wood, S. Goldwater, and M. J. Black. A Non-Parametric Bayesian Approach to Spike Sorting. In *The 28th Annual International Conference on Engineering in Medicine and Biology Society*, pages 1165–1168, 2006.

[Yaglom, 1987] Yaglom. *Correlation Theory of Stationary and Related Random Functions*. Springer-Verlag, 1987.