UNIVERSITY OF CALIFORNIA,
IRVINE

Prediction Constrained Factor Analysis

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Computer Science

by

Madina Abdrakhmanova

Thesis Committee:
Erik Sudderth, Chair
Alex Ihler
Stephan Mandt

2019

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my research advisor and chair of my committee Prof. Erik Sudderth for inspiring this work, being a mentor and guiding me through this process.

I would like to thank my committee members Prof. Alex Ihler and Prof. Stephan Mandt.

Last but not least, I thank my dearest friend and labmate John Gabriel Hope for pushing me through the hardest of times, helping me to focus on the ultimate objectives of the project, advising me on technical issues and supporting me at my thesis defense.

ABSTRACT OF THE THESIS

Prediction Constrained Factor Analysis

By

Madina Abdrakhmanova

Master of Science in Computer Science

University of California, Irvine, 2019

Erik Sudderth, Chair

This work is explores linear dimensionality reduction techniques that preserve information relevant for specific classification tasks. We propose a Gaussian latent variable model that is tuned to maximize the likelihood of the observed data, subject to a constraint that a prediction loss based on the lower dimensional representation meets a chosen threshold. We augment a log-likelihood objective with auxiliary losses that enforce the prediction constraint via Lagrange multipliers. Our prediction-constrained training objective effectively integrates supervisory information even when only a small fraction of training samples are labeled. We analyzed the performance of our PC approach for predicting emotions from face images. We improved prediction quality compared to a multinomial logistic regression model fitted on the output of standard linear dimension reduction techniques. We also achieved competitive performance against a multinomial logistic regression model trained on full resolution images. We were able to learn parameters that can project images to a low-dimensional space and capture the defining feature of each class. In particular, reconstructed images with these parameters were significantly better in capturing the facial expressions compared to Factor Analysis and Probabilistic Principal Component Analysis.

Chapter 1

Introduction

Factor Analysis (FA) is commonly used to extract low-dimensional features that capture the dominant correlations in data. These features are often used as input for pattern recognition algorithms. This unsupervised dimension reduction process is not guaranteed to emphasize or even preserve all of the information necessary for accurate predictions. To address this issue, we must incorporate available labels to find model parameters that produce a suitable projection into an informative low-dimensional space. At the same time, we must take into account the natural asymmetry of prediction tasks. We are interested in predicting labels from data, not the other way around. Thus, we need a projection that does not depend on the availability of labels at test time. This issue is ignored by existing approaches for supervised FA or it's more restricted variation Probabilistic Principal Component Analysis (PPCA) [5]. The problem is exacerbated when we prioritize the generative process of FA and expect to achieve projections with an interpretable structure.

We present Prediction Constrained Factor Analysis, a framework that coherently incorporates the labels into the FA training objective while emphasizing the prediction asymmetry. Our training process prioritizes model parameters that can balance capturing predictive

and generative processes. This training framework was first introduced to topic models by Hughes et al. [2]. In this work, we extend it to derive a training objective for Factor Analysis and show the efficacy of the method on a toy data and a facial image dataset.

Chapter 2

Background

2.1 Supervised Factor Analysis

Factor Analysis is linear Gaussian latent model which assumes that high-dimensional data is a linear function of coordinates on a low-dimensional manifold, plus Gaussian noise[1]. The generative process for each of the observations $x_n \in \mathbb{R}^D$ is described by the graphical model in Fig. 2.1(a) and defined as:

$$
\begin{aligned}
x_n &= Fz_n + \mu + \epsilon \\
z_n &\sim \mathcal{N}(0, I) \\
\epsilon &\sim \mathcal{N}(0, \Psi)
\end{aligned}
$$

where $z_n \in \mathbb{R}^K$ is a latent variable drawn from a Gaussian distribution with zero mean and unit variance. z_n is mapped to a D-dimensional space using the factor loading matrix $F \in \mathbb{R}^{D \times K}$ and the vector $\mu \in \mathbb{R}^D$, that allows a nonzero mean of the data. The noise ϵ is sampled from a Gaussian distribution with zero mean and covariance $\Psi = diag(\psi_1, \psi_2, ..., \psi_D)$.

If we set the constrain $\Psi = \sigma^2 I$, the same variance for each of the dimensions, we get the Probabilistic Principal Component Analysis model (PPCA). Both models assume each observation is conditionally independent given z_n such that:

$$x_n | z_n \sim \mathcal{N}(x_n; Fz_n + \mu, \Psi)$$

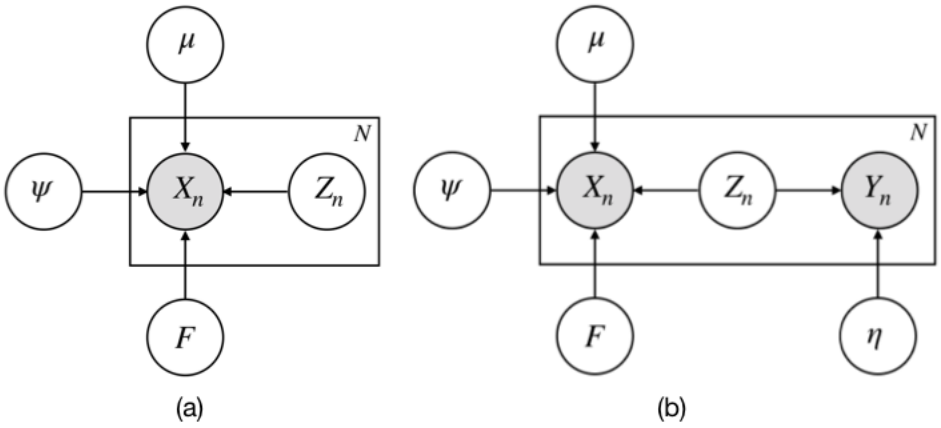

Figure 2.1: (a) Factor Analysis. (b)Supervised Factor Analysis

Suppose each observation x_n has a binary label $y_n \in \{0, 1\}$. Let us assume that the labels are conditionally independent given z_n:

$$y_n | z_n, \eta \sim \text{Bern}(y_n | \sigma(z_n^T \eta))$$

where $\sigma(.)$ is a logit function, and $\eta \in \mathbb{R}^K$ is a vector of regression weights (see Fig. 2.1(b).

2.2 Limitations of existing supervised training objectives

Yu et al. introduced a Supervised PPCA to regression problems [5]. Point estimates of the model parameters are found by maximizing the joint likelihood of observed objects $x_n \in \mathbb{R}^D$ and their labels $y_n \in \mathbb{R}^L$ (see Fig. 2.2(a)). Similarly to x_n, y_n is treated as a continuous random variable with Gaussian distribution. Then, the joint likelihood is defined as

$$p(D|\theta) = \prod_{n=1}^{N} p(x_n, y_n|\theta_x, \theta_y) = \prod_{n=1}^{N} \int p(x_n|z_n, \theta_x) p(y_n|z_n, \theta_y) p(z_n) dz$$
$$p(x_n|z_n, \theta_x) = \mathcal{N}(F_x z_n + \mu_x, \sigma_x^2 I)$$
$$p(y_n|z_n, \theta_y) = \mathcal{N}(F_y z_n + \mu_y, \sigma_y^2 I)$$

where $\theta_x = \{F_x, \sigma_x, \mu_x\}, \theta_y = \{F_y, \sigma_y, \mu_y\}$.

The major problem with this approach is the cardinality mismatch between the observed values. When D, the size of x_n, is much larger than L, the size of y_n, the optimal solution to maximizing the joint likelihood $\max_{\theta_x, \theta_y} p(\mathbf{x}, \mathbf{y}|\theta)$, can be almost identical to the one to maximizing the likelihood of only observed objects: $\max_{\theta_x} p(\mathbf{x}|\theta_x)$. As the result, the impact of the labels in finding the optimal parameters is insignificant and these parameters may not contain the information essential for the regressor.

A simple method to increase the contribution of the low dimensional labels is to replicate each of them. Consider the generative model that was previously introduced for supervised Factor Analysis. Fig. 2.2(b) presents the modified version with label replication. Then, the joint likelihood can be defined as:

$$p(D|F, \mu, \Psi, \eta) = \prod_{n=1}^{N} \int p(x_n|z_n, F, \mu, \Psi) p(y_n|z_n, \eta)^{\lambda} p(z_n) dz$$

Note, the posterior $p(z_n|x_n, F, \mu, \Psi)$, used for making predictions at the test stage, can be different from $p(z_n|x_n, y_{n_1}, y_{n_2}, \dots, y_{n_\lambda}, F, \mu, \Psi)$, which is computed at the training stage. The parameters learned when both $\mathbf{x}$ and $\mathbf{y}$ are observed, can give a latent representation that results in poor performance when no labels are provided. Hughes et al. presented an example where label replication for LDA performed badly when tested on the data it was previously trained on.

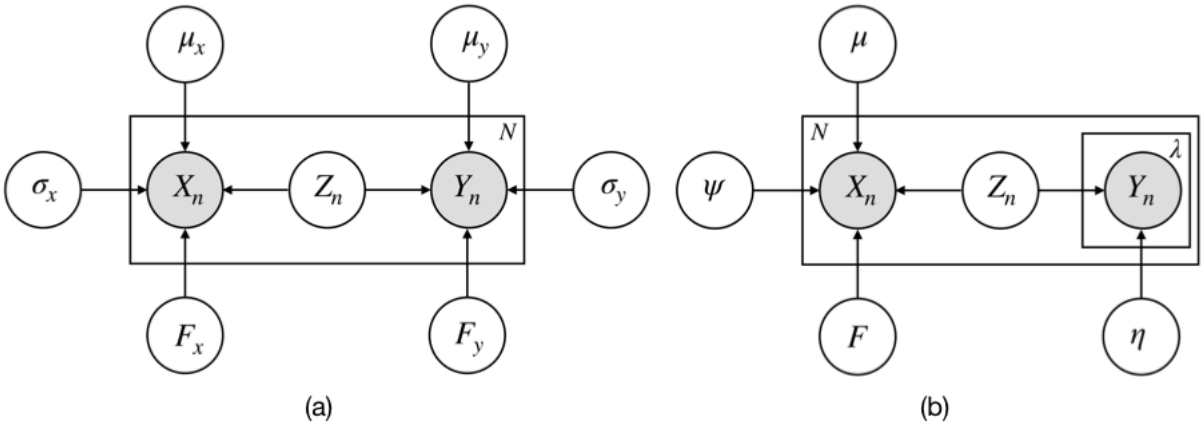

Figure 2.2: (a) Supervised PPCA for regression problems. (b) Supervised Factor Analysis with label replication

Chapter 3

Prediction-Constrained Factor Analysis

In contrast to the methods presented above, we propose a Prediction Constrained (PC) objective that finds the *maximum a posteriori* of parameters F, Ψ, μ, η given a datapoint x_n, while satisfying the constraint that the parameters must yield accurate predictions about labels given x_n alone:

$$\min_{F,\Psi,\mu,\eta} -[\sum_{n=1}^{N} \log p(x_n|F,\psi,\mu)] - \log p(F,\psi,\mu,\eta)$$
$$\text{subject to } [\sum_{n=1}^{N} \log p(y_n|x_n,F,\psi,b,w)] \leq \epsilon$$

where scalar ϵ is the lowest acceptable threshold for the accuracy of our predictions or, equivalently, the highest aggregate loss we are willing to tolerate. $p(F,\psi,\mu,\eta) = p(F)p(\psi)p(\mu)p(\eta)$ are independent priors used for regularization.

Applying Karush-Kuhn-Tucker conditions, we transform the inequality constrained objective

into an equivalent unconstrained optimization problem using a Lagrange multiplier λ_ϵ:

$$\min_{\Psi,\theta,\mu,\eta} -\sum_{n=1}^{N}[\log p(x_n|F,\psi,\mu) - \lambda_\epsilon \log p(y_n|x_n,F,\psi,\mu,\eta)] - \log p(F,\psi,\mu,\eta)$$

where we must search the appropriate value for λ_ϵ over the one-dimensional space and tune for each new data set using cross validation.

From FA model we know:

$$\begin{aligned} p(x_n|F,\psi,\mu) &= \mathcal{N}(\mu, FF^T + \Psi) \\ p(z_n|x_n,F,\psi,\mu) &= \mathcal{N}(x_n|\mu_{z_n|x_n}, \Sigma_{z_n|x_n}) \\ \mu_{z_n|x_n} &= \Sigma_{z_n|x_n} F^T \Psi^{-1}(x_n - \mu) \\ \Sigma_{z_n|x_n} &= \left(I + F^T\Psi^{-1}F\right)^{-1} \end{aligned}$$

As discussed earlier, we assume $y_n \in \{0,1\}$ and $y_n|z_n,\eta \sim Bern(y_n|\sigma(z_n^T\eta))$, where $\sigma(z) = (1+e^{-z})^{-1}$. Computing $\log p(y_n|x_n,F,\psi,b,\eta)$ involves the marginalization of z_n over its domain:

$$p(y_n|x_n,F,\psi,\mu,\eta) = \int_{z_n} p(y_n|z_n,\eta)p(z_n|x_n,F,\Psi,\mu)dz_n$$

Since this integral is intractable, we redefine the objective by instantiating z_n, instead of marginalizing it away. We choose the *maximum a posteriori* (MAP) of z_n given x_n: $z_n = \text{argmax}_z \log p(z_n|x_n)$ which is $\mu_{z_n|x_n}$. Thus, we get a closed form of PC training objective for Factor Analysis:

$$\min_{\Psi,\theta,\mu,\eta} -\sum_{n=1}^{N}[\log p(x_n|F,\psi,\mu) - \lambda \log p(y_n|\mu_{z_n|x_n},\eta)] - \log p(F,\psi,\mu,\eta)$$

We perform automatic differentiation to compute gradients and use Limited-memory BFGS algorithm in Pytorch to find the optimal parameters [[4]].

Note, our approach emphasizes the asymmetric prediction of $y|x$, while the label replication method ignores it and instead emphasizes the connection between the labels and latent variables. In fact, Hughes et al. 2018 illustrated an example how no amount of label replication for standard maximum likelihood training can reach the prediction quality achieved by the PC approach for topic models [2].

Chapter 4

Results and Discussion

4.1 Experiments description

We assess the performance of PC training of FA on two datasets:

1. The toy data represents points drawn from a pair of two-dimensional Gaussian distributions such that the direction of maximum variance is parallel to the optimal linear decision boundary. We project the data onto one dimensial space.

2. The AR Face database contains nearly 1500 facial images taken from the frontal view of 126 human subjects (70 male and 56 female). The images capture four emotions(neutral, smile, angry and scream) and contain two additional illumination conditions for the neutral mode (left and right light). We combined neutral, left light and right light into one class, and with the rest of the three categories focused on classifying the four emotions. The images were aligned and cropped to size 150×150 pixels (see Fig. 4.1)[3].

The two baselines we consider are logistic regression fitted on raw pixel data and logistic

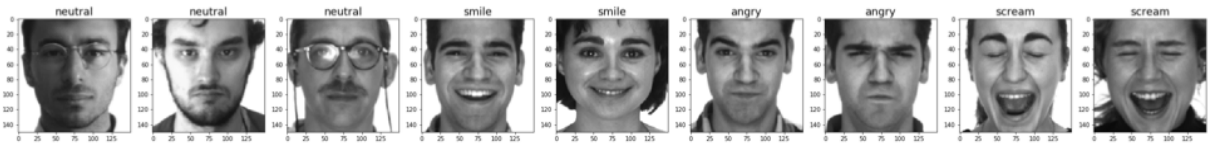

Figure 4.1: Aligned and cropped images from AR Face Database for each of the 6 categories: neutral with left light, neutral with right light, neutral, smile, angry and scream.

regression fitted on the transformed data by standard FA or PCA. For each of the classification tasks we explored the effect varying λ_ϵ on the performance of the model. We also examined the role of regularization in learning the parameters.

4.2 Toy data

Since our toy data is homogeneous, we used PPCA version of our model where the noise variance is constrained as $\Psi = \sigma^2 I$. Fig. 4.2(b) presents a plot of the toy data, where the color of the samples corresponds to their true class label. The green line is the axis of projection produced by PCA from sklearn package implementation. The red and purple lines are the axes of projection produced by PC PPCA with $\lambda_\epsilon = 0$ and $\lambda_\epsilon = 2$ respectively. When $\lambda_\epsilon = 0$, the PC model is equivalent to standard PPCA. The figure shows that the PC PPCA solution found through gradient descend optimization indeed matches the closed

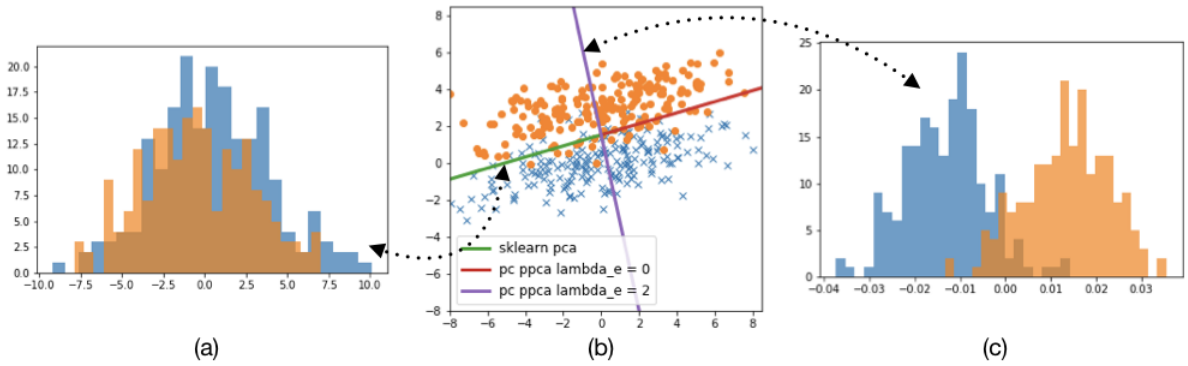

Figure 4.2: (a) The histogram of samples transformed by PCA. (b) The 2D toy data and the axis of projections produced by standard PCA and PC PPCA with varying λ_ϵ (c) The histogram of samples transformed by PC PPCA with $\lambda_\epsilon = 2$

form solution computed by the standard PCA. Fig. 4.2(a) illustrates the transformation of samples by PCA. These data points are colored in accordance with their true label as it was done in Fig. 4.2(b). The histogram shows that the resulting projection will only further confuse the logistic regression classifier fitted onto it. Fig. 4.2(c) presents the effect of the higher prediction weight on transformation of the samples. That is, emphasizing the asymmetric prediction of $y|x$, indeed helps to find the latent space that is useful for the logistic regression classifier.

Fig. 4.3 shows the performance of PC training for PPCA over a range of values for λ_ϵ. As the prediction weight value gradually increases from 0 to 2, the training accuracy increases, the slope of the produced projection axis comes closer to the optimal solution illustrated in Fig. 4.2(b), and reconstruction error increases in sacrifice for the improved accuracy. When we have no regularization on the model parameters, we observe how an infinitesimal step in the value of the prediction weight gives a drastic change in all of the three metrics. As we are increasing regularization, the change in the metric values becomes smoother.

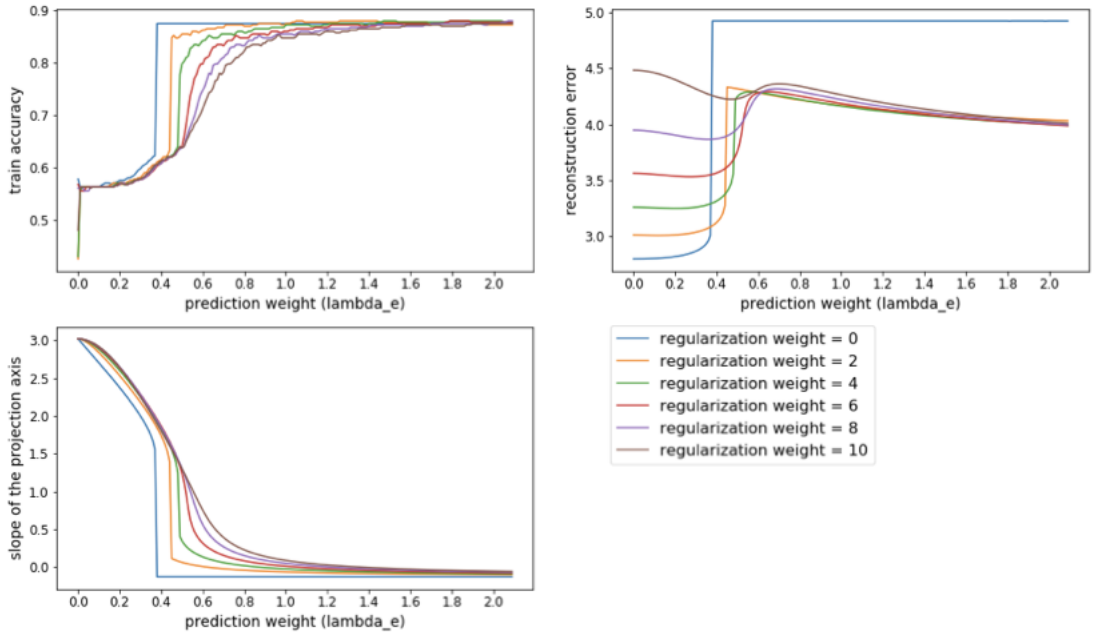

Figure 4.3: The effect of regularization on the performance of PC PPCA with varying λ_ϵ

4.3 AR data

Baseline results: We examined the performance of multinomial logistic regression fitted on the data points transformed by standard PCA or FA with a varying number of components and compared them with the result of the baseline, the multinomial logistic regression fitted on original images. The baseline achieves 97.7% and 86.08% accuracy on train and test sets respectively. As the confusion matrices in Fig. 4.4 show, the majority of errors are due to the misclassifications between "angry" and "neutral" classes, which can be explained by the natural similarity between the two expressions.

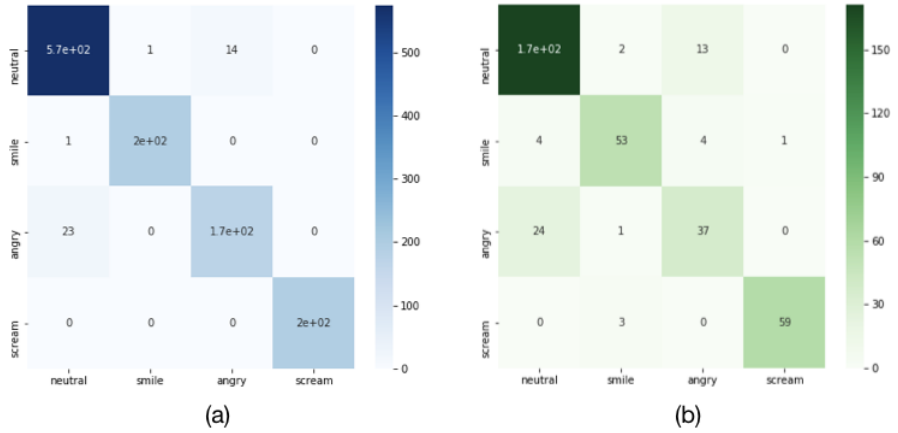

Figure 4.4: Multinomial logistic regression fitted on the images of size 150 × 150 pixels (a) Confusion matrix on train set. (b) Confusion matrix on test set

Fig. 4.5 shows that as the number of components becomes higher, both model variations are getting closer to the baseline results. The reconstruction error decreases since we can preserve a higher percentage of original variance with a larger number of components. Comparing to PPCA, FA consistently produces better projections to low dimensional spaces since it has a greater degree of freedom in noise variance. Therefore, we focused on PC training in the latent dimensions that provide a substantial room for improvement, such as 3, 10 and 20. The plot on test accuracy depicts the results of the best PC models in these dimensions. In latent dimension 10, both model settings can already achieve scores comparable with the

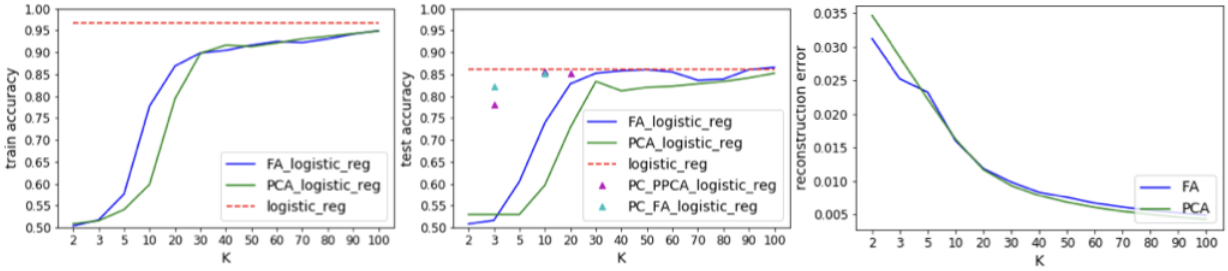

Figure 4.5: The performance of PCA and FA on AR data with varying number of components K

baseline result. In latent dimension 3, PC FA improves the score of PC PPCA by nearly 5%.

PC PPCA results: We analyzed the performance of the parameters learned through PC training by projecting the full resolution images to a low-dimensional space (3, 10, 20) and then reconstructing them in accordance with the generative process of Factor Analysis. Fig. 4.6 compares the quality of reconstruction from latent spaces with dimension 10 produced by PCA and PC PPCA with $\lambda_\epsilon = 1, 1000, 10000$. Fig. 4.6(b) illustrates how PCA is able to capture only the most obvious variance in the pixel values caused by the varying illumination conditions in the neutral mode. Most importantly, the model completely ignores the defining characteristics of each emotion. When $\lambda_\epsilon = 1$, the PC objective is equivalent to the standard joint maximum likelihood. The resulting projection leads to a slight improvement in test accuracy score comparing to PCA(59.69% vs 64.25%), but the reconstructions in Fig. 4.6(c) show that the prediction weight is too small for the labels to make a significant difference. In contrast, prediction weight values 1000 and 10000 are high enough to emphasize the variance that characterizes smiling and screaming emotions, achieve test accuracies comparable with the baseline, 84.68% and 85.75% respectively, while keeping the noise level low.

When we increase the dimension of the latent space to 20, we can now capture some subject specific details (see Fig. 4.7). Overall, the presented models behave similarly to the corresponding ones with latent dim 10. When we project to a 3-dimensional latent space,

Figure 4.6: (a) Original images from AR Face database. The images are projected to a 10-dimensional latent space. (b) Image reconstruction from PCA. (c) Image reconstruction from PC PPCA with $\lambda_\epsilon = 1$. (d) Image reconstruction from PC PPCA with $\lambda_\epsilon = 1e3$. (e) Image reconstruction from PC PPCA with $\lambda_\epsilon = 1e4$

.

Figure 4.7: (a) Original images from AR Face database. The images are projected to a 20-dimensional latent space. (b) Image reconstruction from PCA. (c) Image reconstruction from PC PPCA with $\lambda_\epsilon = 1$. (d) Image reconstruction from PC PPCA with $\lambda_\epsilon = 1e3$. (e) Image reconstruction from PC PPCA with $\lambda_\epsilon = 1e4$

the best PC model configuration with $\lambda_\epsilon = 10000$ is able to improve on the test score based on PCA projection to 77.95%, but the reconstructions lack the quality level of their higher dimensional counterparts (Fig. 4.8).

Figure 4.8: (a) Original images from AR Face database. The images are projected to a 3-dimensional latent space. (b) Image reconstruction from PCA. (c) Image reconstruction from PC PPCA with $\lambda_\epsilon = 1$. (d) Image reconstruction from PC PPCA with $\lambda_\epsilon = 1e3$. (e) Image reconstruction from PC PPCA with $\lambda_\epsilon = 1e4$

PC FA results: Since FA has a higher degree of freedom in noise variance Ψ, PC FA with $\lambda_\epsilon = 1e4$ is able to achieve a higher test score of 82.26% than PC PPCA with the same prediction weight. It also finds a much better projection than standard FA as illustrated in Fig. 4.9. The right column of the figure presents 3D projection by FA. Similarly to PCA, FA emphasizes the variance due to the three illumination conditions, and groups the data points accordingly. The dark blue and mid blue correspond to right and left illumination modes respectively. The third group of points are the transformed images taken under normal conditions. As the result the projection does not hold any useful information for a classifier which is reflected in low test score of multinomial logistic regression 51.61%. For example, the bottom left plot indicates that the classifier mislabels all of the "smile" examples. In

contrast, PC FA arranges the 3D points in accordance with not only their lightning modes but also their true class labels(Fig. 4.9 (c)). For instance, there are distinct clusters of red "scream" and green "smile" points. The mix of orange "angry" and light blue "neutral, no illumination" points corresponds to the confusion between these two classes as it happened for multinomial logistic regression fitted on high dimensional images. It could be explained by the same reason of similarity between these expressions among test subjects. Such projection is much more suitable for a classifier and results in comparable to baseline accuracy.

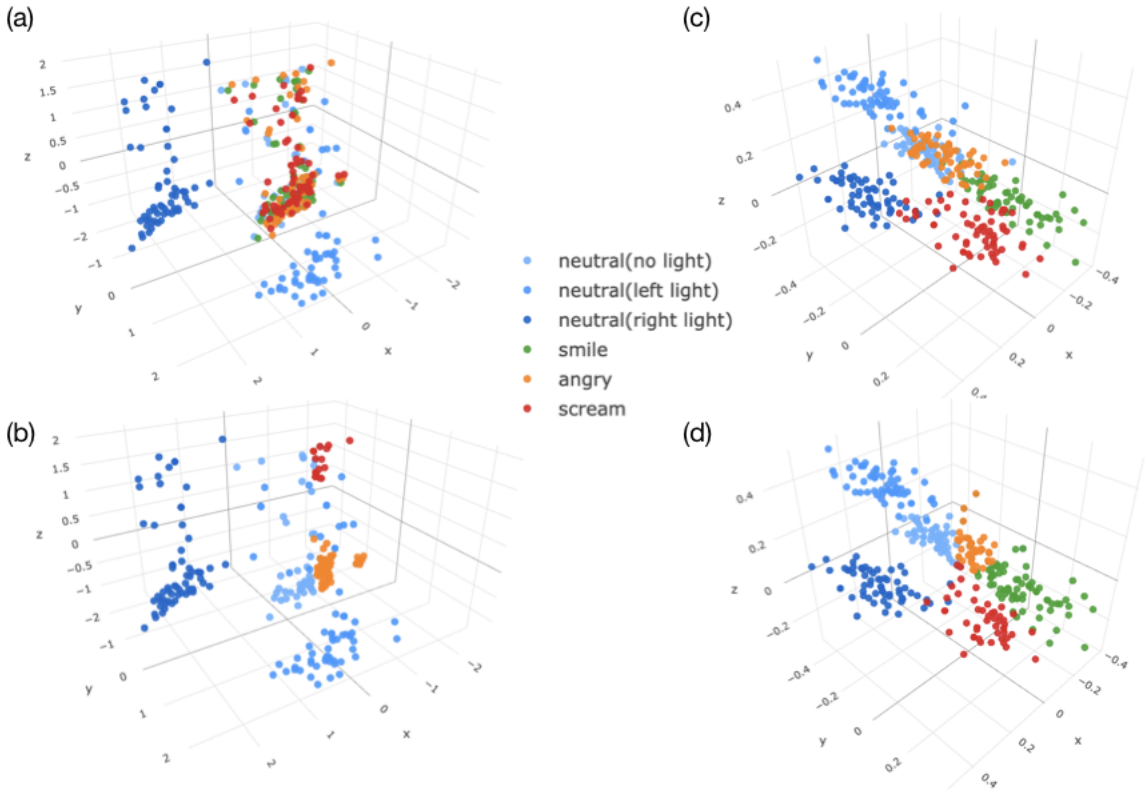

Figure 4.9: The images from AR Face database are projected to a 3-dimensional latent space. The left side: the test samples are transformed using FA and colored in accordance with (a) the true labels and (b) the predicted labels by multinomial logistic regression. The right side: the test samples are transformed using PC FA with $\lambda_\epsilon = 1e4$ and colored in accordance with (c) the true labels and (d) the predicted labels by multinomial logistic regression whose weights were learned alongside with the rest of the parameters.

Allowing a different variance for each of the original dimensions is indeed helpful in finding a suitable low-dimensional projection for a downstream prediction task. However, this can

also lead for some entries in Ψ to be exceptionally high, which can cause significant noise in reconstructions. In particular, reducing the dimension of gray scale images to 10 with $\lambda_\epsilon = 10000$ gives high accuracy but very noisy reconstructions (see Fig.4.10(c)). Introducing a log-normal prior with appropriate mean and covariance for Ψ, helps to denoise the reconstructions and slightly improve the accuracy (see Fig.4.10(d)).

Figure 4.10: (a) Original images from AR Face database. The images are projected to a 10-dimensional latent space. (b) Image reconstruction from FA. (c) Image reconstruction from PC FA with $\lambda_\epsilon = 1e4$. No regularization on noise variance. (d) Image reconstruction from PC PPCA with $\lambda_\epsilon = 1e4$. With regularization on noise variance.

Chapter 5

Conclusion

In this work, we extended the Prediction Constrained training framework to Factor Analysis. Comparing to the alternative supervised approaches, we explicitly incorporated the asymmetry of predicting labels from observed data. Thus, our method was able to find a low-dimensional manifold that provides accurate predictions and sheds light on how they are made. We demonstrated the performance of PC FA on a toy binary classification task and an emotion prediction task based on the AR face database. We found that the parameters learned through PC training gave interpretable low-dimensional projections that were useful for our target classification tasks. The major limitation of our approach lies in the assumption of FA that high-dimensional data lies close to a linear subspace. Thus, out method works well with images that are perfectly aligned and cropped, but would fail on more realistic facial image data sets. To address these issues in future work, I would like to extend PC training approach to variational autoencoders. These more flexible models are parametrized by neutral network and can capture non-linear dependencies.

Bibliography

[1] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, New York, NY, USA, 2012.

[2] M. Hughes, G. Hope, L. Weiner, T. McCoy, R. Perlis, E. Sudderth, and F. Doshi-Velez. Semi-supervised prediction-constrained topic models. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1067–1076, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[3] A. M. Martinez and R. Benavente. The ar face database. *Tech. Rep. 24 CVC Technical Report*, 01 1998.

[4] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

[5] J. Xia, J. Chanussot, P. Du, and X. He. (semi-) supervised probabilistic principal component analysis for hyperspectral remote sensing image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2224–2236, June 2014.