Using the Trace Norm Prior to Extend Mixture of Subspace Models

Jason D. M. Rennie
jrennie@gmail.com

June 20, 2006*

1 Mixture Models

A popular approach for both classification and clustering is to assume that data is generated by a mixture model. Support of the mixture model for classification has waned as discriminative models have gained popularity. However, the mixture model is a valuable element of statistics and remains useful for clustering. Furthermore, reasoning about mixture models is often relatively simple, so they may still provide valuable intuition for the development of classification models.

A mixture model assumes that each datum is generated via a two-stage process. First, a class is selected according to a multinomial distribution. Second, a datum is generated for the selected class. Typically, class models have a common form, but different parameter settings.

In this work, we will discuss mixture models where the individual class models generated data in a low-dimensional subspace. This type of model is sometimes called a "mixture of subspaces" model.

2 Previous Work

Hinton et al. [3] discuss two mixture of subspaces models. The first uses Principal Components Analysis (PCA) for the class model. The second uses a Factor Analysis (FA) class model.

2.1 Principal Components Analysis

Given data $X \in \mathbb{R}^{n \times d}$, PCA finds the $k < d$ orthogonal directions of maximal variance. In other words, PCA finds

$$(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \arg \max_{\mathbf{u}_1 \perp \cdots \perp \mathbf{u}_k, \|\mathbf{u}_i\|=1 \forall i} \sum_{i=1}^{k} \|X\mathbf{u}_i\|_2^2. \tag{1}$$

*Updated June 29, 2006.

The orthonormal vectors $(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ are known as the principal components and define a subspace of $\mathbb{R}^d$. Viewed as a likelihood model, PCA is an improper Gaussian, improper since it cannot be normalized. PCA uses infinite variance in the directions of the principal components and unit variance elsewhere. Let V be an orthogonal matrix where the first k rows are the principal components. Then the PCA likelihood distribution is $\mathcal{N}(\mathbf{0}, V\Sigma V^T)$, where Σ is diagonal with first k entries ∞, and remaining entries unity, $\Sigma = \text{diag}(\infty, \ldots, \infty, 1, \ldots, 1)$,

$$P_{\text{PCA}}(X|\mathbf{v}_1, \ldots, \mathbf{v}_k) \propto \exp(-XV\Sigma^{-1}V^T X^T) \tag{2}$$

As Hinton et al. note, the PCA "model" is seriously deficient. Viewing the classification/clustering task as one of transmitting labels, the PCA model ignores the costs of communicating (1) the model, and (2) the projections of the data onto the principal components. It is also rigid in that it uses a constant variance for the non-principal component directions.

2.2 Factor Analysis

Factor Analysis (FA) can be viewed as an extension to PCA which fixes two of the three above-mentioned issues. A set of k "factors" (principal components) are chosen by the model. Whereas PCA simply ignores the factor "space," FA provides a full, normalizable model. The data is assumed to be a sum of the factors projected into data space, plus noise,

$$\mathbf{x} = \Lambda\mathbf{z} + \mathbf{u}, \tag{3}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_k)$, and $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \text{diag}(\boldsymbol{\psi}))$ [2]. I.e. the components of the data are independent given the factors. The interpretation of the factors are different from that of the principal components. Whereas PCA finds directions of maximum variance, FA has parameters to specifically capture axis-aligned variance ($\boldsymbol{\psi}$). The factors are used to allow additional control over a k-dimensional subspace. I.e. the underlying data is assumed to be a transformation from a k-dimensional subspace; the observed data is corrupted with simple Gaussian noise. The FA likelihood is

$$P_{\text{FA}} \propto \exp(-X[\Lambda\Lambda^T + \text{diag}(\psi)]^{-1}X^T). \tag{4}$$

As long as the noise in all dimensions is non-zero, $\boldsymbol{\psi} > 0$, the covariance matrix $(\Lambda\Lambda^T + \text{diag}(\psi))$ is positive definite and thus the data likelihood is normalizable.

3 Modeling the Factors

Though the factors ($\mathbf{z}$) are technically modeled by a unit Normal distribution, they are, in effect, only restricted in number. The "factor loading matrix" (Λ) yields an arbitrary rank k covariance matrix for the distribution of the factors in data space, $\Lambda\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \Lambda\Lambda^T)$. Hence, except for the rank constraint, which limits the number of factors, and the zero-mean Gaussianity assumption, there is no preference for how the factors are distributed—there is no preference in the choice of factor loading matrix. In other words, Factor Analysis (FA) does not account for the cost of communicating the model (the directions of the factors in data space).

Here we extend FA to account for the cost of the model. We do this by introducing a prior on the factor loading matrix. This may sound simple, but is, in fact, subtle. The idea behind Factor Analysis is to presume that underlying data comes from a low-dimensional sub-space. However, a simple matrix prior, such as the Frobenius norm[1] does not encourage a low-rank solution, much like a Gaussian prior or L_2-norm penalty does not encourage zero weights. What we would like is a prior which is a generalization of the Laplacian prior or L_1-norm weight penalty. The trace norm[2], or sum of singular values of a matrix is one such generalization. In fact, it is likely the only generalization which is not limited in the type of (real) matrices to which it can be applied. The trace norm can be trivially extended to a concave prior over matrices which encourages low rank,

$$P(X) \propto \exp(-\lambda \|X\|_\Sigma) \tag{5}$$

[4], where $\lambda > 0$ is a constant which controls the strength of the prior. λ can be thought of as a continuous version of the rank, k. However, the relationship is inverted. A small value of λ encourages a high-rank solution, whereas a large value encourages a low-rank solution.

Our extension of Factor Analysis to include a modeling of the factors is to simply include the trace norm distribution as a prior on the factor loading matrix. Whereas before the factor loading matrix was rank-limited, here we allow it to be full-size, $\Lambda \in \mathbb{R}^{d \times d}$. Instead of imposing a hard constraint on the rank of the factor loading matrix, we use the trace norm prior as a soft penalty which encourages low rank.

Let $S \equiv \Lambda^T \Lambda + \mathrm{diag}(\boldsymbol{\psi})$. Then, the updated model is

$$P(X|\Lambda, \boldsymbol{\psi}) P(\Lambda|\lambda) \propto \frac{\exp\left(-\frac{1}{2}\mathbf{Tr}\left[X S^{-1} X^T\right]\right)}{|2\pi S|^{1/2}} \exp(-\lambda \|\Lambda\|_\Sigma), \tag{6}$$

The model is still incomplete, as we have yet to add (informative) priors for $\boldsymbol{\psi}$ and λ. However, we have substantially reduced the number of parameters without a prior. And, additional priors might make comparison of the two models more difficult.

Note that our updated FA model with a trace norm prior may be easily substituted for the FA model in Mixtures of Factor Analyzers clustering framework (see e.g. §3 of [3]).

[1] The Frobenius norm of a matrix, X, is the square root of the sum of the squared entries, $\|X\|_{\mathrm{Fro}}^2 = \sum_{i,j} X_{i,j}^2$.

[2] The trace norm is also known as the nuclear norm and the Ky-Fan norm. See Fazel [1] §5 for a development of the trace norm and discussion of some of its properties.

References

[1] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

[2] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Department of Computer Science, University of Toronto, 1996.

[3] G. E. Hinton, P. Dayan, and M. Revow. Modeling the manifolds of images of handwritten digits. *IEEE Transactions on Neural Networks*, 8(1):65–74, 1997.

[4] J. D. M. Rennie. Text modeling with the trace norm. http://people.csail.mit.edu/jrennie/writing, April 2006.