REVIEW ON IEEE TRANSACTION ON ML AND MCE FACTOR ANALYSIS FOR AUTOMATIC SPEECH RECOGNITION

Naveen Parihar

Mississippi State University
Mississippi State, MS 39762 USA
email: parihar@isip.mstate.edu

ABSTRACT

The paper claims that the combined use of mixture densities and factor analysis for speech recognition leads to smaller, faster and accurate recognizers than either of these in isolation. Two ways to model correlations between high-dimensional feature vectors are: 1) implicitly by the use of mixtures, or 2) explicitly by the use of non diagonal elements in each covariance matrix. The latter one has a heavy computational overhead because of the use of full covariance matrix. Factor analysis can be used to model this high dimensional covariance matrix using small number of parameters. Factor analysis is not only a method for dimensional reduction but it also models the variations outside the reduced-dimensionality subspace. Factor analysis can be used to increase likelihoods as well as word accuracies by use of an expectation-maximization (EM) algorithm for maximum likelihood estimation and a gradient descent algorithm for improved class discrimination. This paper will analyze the use of factor analysis technique in conjunction with mixture densities to model correlations by reviewing advantages and drawbacks of the EM algorithm and gradient descent algorithm.

1. INTRODUCTION

In a speech recognizer, the spectral information within a frame is represented by a feature vector consisting of approximately 30 dimensions. Correlations between these features may exists when the speech signal is either non stationary or is corrupted by noise. Background noise and coarticulation effects give rise to continuos variability. Currently, many Hidden Markov Models which are used to model the short-time, acoustic properties of speech, ignore this correlation. Though, by the use of two or more mixtures of Gaussian probability density functions (PDF's) with diagonal covariances matrices, these correlations are modelled implicitly, there is no viable explicit technique to model these correlations. The use of full covariance matrices is not a viable solution to model these correlations explicitly since it involves a very large computational overhead. and it is very difficult to get the full covariance matrices because of lack of enough data. The other technique of sharing the covariance matrices across the states or models, parameter-tying, is very complicated. The paper claims that the statistical method of factor analysis is a compromise between the two extreme approaches of the use of either full covariances or diagonal matrices to model correlations. Factor analysis is a linear technique of dimensional reduction. In other words, factor analysis maps the high dimensional space into a lower dimensional subspace by expressing the full covariance matrices in terms of small number of parameters. In this way, factor analysis captures the most significant correlations. Thus, it has a very small overhead in computation and memory requirements in comparison to the use of full matrices to model correlations explicitly.

The paper also claims that the use of mixture densities to model discrete type of variability and the use of factor analysis to model the continuos type of variability are two complementary techniques. The combined use of both in Hidden Markov Model leads to smaller, faster and accurate recognizers than either of these in isolation. While the former can be understood as clustering, the latter is dimensional reduction technique.

The small number of parameters of factor analysis can be chosen in two ways either to increase likelihoods or to improve word accuracies by use of

an expectation-maximization (EM) algorithm for maximum likelihood estimation or by use of a gradient descent algorithm for improved class discrimination.

2. FACTOR ANALYSIS

We will discuss dimensionality reduction by using factor analysis, learning algorithm for ML and MCE using factor analysis in sections 2.1, 2.2 and 2.3, respectively. For each of these sections we will first consider the analysis for multivariate Gaussian PDF's and then extend the method to HMM's using multiple mixtures per state. Here each mixture represents a Gaussian PDF.

2.1. DIMENSIONALITY REDUCTION

The objective of the factor analysis is to reduce the dimensions that captures the correlations among the features.

If $x \in R^D$ denotes a Gaussian random variable with mean μ, then the number of dimensions D is very large. Factor analysis reduces it to $f \ll D$ dimensions. Here $z \in R^f$ is also a Gaussian random variable with zero mean and identity covariance matrix. The marginal distribution of x is given by

$$P(x) = \frac{\left|\Lambda^T\Lambda + \psi\right|^{-1/2}}{(2\Pi)^{D/2}} \exp\left\{-\frac{1}{2}[x-\mu]^T(\psi + \Lambda\Lambda^T)^{-1}[x-\mu]\right\} \quad (1)$$

Here Λ denotes an arbitrary $D \times f$ *factor loading* matrix and ψ denotes a diagonal $D \times D$ matrix.

The variances ψ_{ii} denotes the variances outside the subsample $S(\Lambda)$ and maximum variation in x is captured by the columns of Λ. Clearly, since $f \ll D$, storing Λ and ψ requires less memory than storing a full covariance matrix. Also, the covariance matrices of these form can be inverted using matrix inversion lemma[2]. Using which $P(x)$ can be computed with only $O(fD)$ multiplies instead of $O(D^2)$ multiples when a full covariance matrix is used.

The HMM's use the mixtures of these PDF's to implicitly model correlations. The factor analysis HMM or FM-HMM's has mixtures modelled by the PDF's given above. Here distribution for each state is estimated by

$$P(x|s) = \sum_c P(c|s)P(x|s,c) \quad (2)$$

Since, we employ PDF's which requires less number of computations and memory to estimate the correlations, the $P(x|s)$ can be computed in fewer number of computations than required for a full covariance matrix.

2.2. ML FACTOR ANALYSIS

Maximum likelihood criterion is used to estimate parameters of the HMM's. The EM algorithm is a iterative process to estimate the parameters of latent variable models. If we have N data points, then the EM algorithm is a two step iterative procedure to estimate the parameters Λ, ψ and μ that maximize the log likelihood. The first or the E-step is to calculate the Q-function given by

$$Q(\tilde{\mu}, \tilde{\Lambda}, \tilde{\psi}; \mu, \Lambda, \psi) = \sum_n \int dz P(z|x_n, \mu, \Lambda, \psi) \ln P(z, x_n|\tilde{\mu}, \tilde{\Lambda}, \tilde{\psi}) \quad (3)$$

The second or the M-step is to maximize the Q-function using the following iterative step

$$\tilde{\Lambda} \leftarrow \left(\sum_n (\Delta x_n)(\Delta z_n)^T\right) \left(\sum_n [E[\delta z \delta z^T|x_n] + (\Delta z_n)(\Delta z_n)^T]\right)^T \quad (4)$$

(5)

$$\tilde{\mu} \leftarrow \frac{1}{N}\sum_n (x_n - \tilde{\Lambda}E[z|x_n]) \quad (6)$$

$$\tilde{\psi}_{ii} \leftarrow \frac{1}{N}\sum_n [(\Delta x_n - \tilde{\Lambda}\Delta z_n)_i^2 + (\tilde{\Lambda}E[\delta z \delta z^T|x_n]\tilde{\Lambda}^T)_{ii}] \quad (7)$$

These updates are guaranteed to converge monotonically to a extremum of the log-likelihood.

These parameters are estimated using the assumption that the likelihood is unimodal and approximately symmetric. But the likelihood can be multimodal and asymmetric, which undermines the maximum likelihood criterion.[3]

The same iterative process can be used to estimate the parameters of each mixture in every state of the HMM's, except that the each observation is weighted by the posterior probability given by the following equations.

.

$$\tilde{\Lambda}_{sc} \leftarrow \left(\sum_n \gamma_n^{sc}(\Delta x_n^{sc})(\Delta z_n^{sc})^T\right) \left(\sum_n \gamma_n^{sc}[E_{sc}[\delta z \delta z^T | x_n] + (\Delta z_n^{sc})(\Delta z_n^{sc})^T]\right)^T \quad (8)$$

$$\tilde{\mu}_{sc} \leftarrow \frac{1}{N^{sc}} \sum_n \gamma_n^{sc}(x_n - \tilde{\Lambda}_{sc} E_{sc}[z|x_n]) \quad (9)$$

$$[\tilde{\psi}_{sc}]_{ii} \leftarrow \frac{1}{N^{sc}} \sum_n \gamma_n^{sc}([(\Delta x_n^{sc} - \tilde{\Lambda}_{sc}\Delta z_n^{sc})_i^2] \; plus(\tilde{\Lambda}_{sc} E_{sc}[\delta z \delta z^T | x_n]\tilde{\Lambda}_{sc}^T)] \quad (10)$$

The updates for mixture weights and transition matrices in FA-HMM's are similar to the conventional full covariance matrix HMM since the FA-HMM's are form the subset of HMM's whose mixture employ full covariance matrix.

2.3. MCE FACTOR ANALYSIS

The maximum likelihood criterion is not always the desired one because though it does maximizes the likelihood, it does not guarantees minimum error rate. We cannot use MAP decision for minimum classification error criterion since one cannot estimate the parameters of the true distribution and the finite training set [4]. The goal of MCE training approach is correctly discriminate data rather than to fit the distributions to the data as done in maximum likelihood. The error rate estimated from the limited training data is a piece wise constant function of the classifier parameter, it is not differentiable functions of log-likelihoods. Using the smoothed MCE loss function J [4(23)], we can update the parameters iteratively by gradient descent.

$$J = \frac{1}{N}\sum_n \frac{1}{1 + \exp(-d(x_n))} \quad (11)$$

$$\Phi \leftarrow \Phi - \eta(\Phi)\frac{\partial J}{\partial \Phi} \quad (12)$$

where $\eta(\Phi)$ is a positive learning rate.

The partial derivative in (12) can be decomposed using the log-likelihoods [4(25)].

$$\frac{\partial J}{\partial \Phi} = \sum_y \frac{\partial J}{\partial \ln P(x|y)} \frac{\partial \ln P(x|y)}{\partial \Phi} \quad (13)$$

For gradients of ln. $P(x|y)$ with respect to μ_i, Λ_{ij}, ψ_{ij} are given by [1(27-29]. Using these equations iteratively, the variance parameters are updated in the log domain by choosing the positive matrix $\eta(\Phi)$.

The same iterative process is used to estimate each mixture component in a given state of HMM's.

3. EXPERIMENTS

The experiments were conducted on *Connected Alpha-Digits* (36 words) and *New Jersey Town Names (1219 words)*. Various combinations of number of mixture components and number of factors in each state of HMM were evaluated. This is a reasonable way to find out the performance of the system at different Factors.

For *Connected Alpha*-Digits, the results in Table I[1] showed that there is a significant increase performance by increasing the number of factors.

The graphs between average likelihood versus the number of parameters and word error rate versus the number of parameters clearly shows that models with factored covariance matrices have better likelihood and word error rates than models with only diagonal covariances matrices. Table II[1], shows that varying the number of factors in each component give better word error and log-likelihood in comparison to a fixed number of factors in each component but the CPU time also increases. In other words, the performance does not increase. But there is an improvement of about a factor of two in speed and memory over diagonal covariance matrix HMM's with the same accuracy levels.

For *New Jersey Town Names,* only factored HMM's with one factor per mixture component were employed since in previous experiment largest improvement (per parameter) occurred with small number of factors. We observe that the rate of improvement in performance is not linear with the increase in the number of factors per mixture components though it is comparable to the diagonal covariance matrix HMM's. Form the figure 3[1], it is evident that the word error rate increases and then drops back with increase in number of parameters in MCE-based factor analysis, this increase in WER is contradictory to the view that the performance increases using factor analysis.

4. SUMMARY

The paper showed that the factor analysis is not only a method for dimensional reduction but it also models the variations outside the reduced-dimensionality subspace. The variance matrix ψ models the variations outside the subspace. Factor analysis can be used to increase likelihoods as well as word accuracies by use of an expectation-maximization (EM) algorithm for maximum likelihood estimation and a gradient descent algorithm for improved class discrimination. It is also established in the paper that the number of computations for factored analysis are very less than the number of computations required for full covariance matrices analysis.Thus, the factor analysis for automatic speech recognition lies between the two extreme techniques of full covariance matrix and diagonal covariance technique analysis.

The experiments conducted very well support the above mentioned conclusions. The experiments conducted, establish the fact that for the same accuracy, the factored analysis HMM's are faster and smaller than diagonal covariance matrix HMM's. In the MCE based task for *New Jersey Town Names,* the error rate showed an increase with the increase in parameters which contradicts the factored MCE analysis for minimum error rate.

Factor analysis and mixture components were used as a complementary techniques to improve the performance of the system. The system performed better by employing both the techniques rather than either of these. Since performance is improved by tying full covariance matrices, the paper suggests that the clever tying of factor loading matrices across units, states, and/or mixture components would lead to further improvement in the performance. This promising idea may be further explored. Also, the method of factor analysis may be applied to arbitrary features that model short time properties of speech.

REFERENCES

[1] L. K. Saul and M. G. Rahim, "Maximum Likelihood and Minimum Classification Error Factor Analysis for Automatic Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 115–125, March 2000.

[2] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, Numerical *Recipes in C: The Art of Scientific Computing.* Cambridge Univ. Press 1992.

[3] D. Rubin and D. Thayer, "EM algorithms for factor analysis," *Psychometrika,* vol. 47, pp. 69-76, 1982.

[4] B. H. Juang, W. Chou, and C. H. Lee, "Minimum classifiation error rate methods for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 266-277, 1997.

[5] R. O Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2000.