ABSTRACT

Modeling data with Gaussian distributions is an important statistical problem. To obtain robust models one imposes constraints on means and covariances of the data. Constrained ML modeling implies the existence of optimal feature spaces where the constraints are more valid [2, 3]. This paper introduces one such constrained modeling technique called factor analysis, invariant to linear transformations (FACT) which is essentially factor analysis in optimal feature spaces. FACT is a generalization of several existing methods for modeling covariances. This paper presents an EM algorithm for FACT modeling.

1. INTRODUCTION

In Gaussian Modeling the model parameters (means and covariances) are usually estimated using the Maximum Likelihood (ML) principle. In many applications due to data insufficiency, computational and/or storage considerations, one has to constrain the means and covariances so that there are fewer parameters to estimate (e.g., diagonal covariances, reduced-rank means, shared covariances etc.). In such cases it is desirable to model in a feature space in which the constraints are maximally satisfied. This paper introduces one such constrained modeling technique called factor analysis, invariant to linear transformations (FACT). FACT is a direct generalization of standard factor analysis and semi-tied covariance modeling. This paper presents an EM algorithm for FACT parameter estimation and gives results of speech recognition experiments on two large vocabulary continuous speech recognition tasks: the first a dictation task on an IBM internal database and the second on a voicemail database distributed by LDC.

2. DATA MODELING

The basic problem considered here is obtaining a statistical model of observed data where each sample is independent of any other and drawn from a finite set S of Gaussian distributions. Each sample is associated with a single Gaussian or state s ∈ S viz., the observations are conditionally independent given the state sequence. If $A \in \mathbb{R}^{k \times d}$, and $s_1, s_2, \ldots, s_T = s$ are the observation and state sequences, this corresponds to the following model for the data:

$$p(x_t, s_t) = p(s)p(x_t|s_t) = p(s) \prod_{t=1}^{T} p(x_t|s_t).$$

Such models occur often in practice. For example, if the state sequence is iid this corresponds to a Gaussian mixture model of the data; if the state sequence is Markovian this corresponds to a Hidden Markov Model (HMM) of the data with Gaussian observations. This description also includes HMMs with Gaussian mixture observations if the underlying state-sequence is the Gaussian mixture component sequence. Whatever be the underlying p(s), the p(x_t|s) is then completely described by state means (μ_i) and covariances (Σ_i). In practical applications the estimation of the means and covariances are constrained. Well-known examples are Linear Discriminant Analysis (where means are constrained to lie in a lower dimensional space $\text{Span}(\mu_i) = k \leq d$) and Diagonal Covariance (DC) Modeling where covariances are constrained to be diagonal viz., $\Sigma_i = D_i, D_i$ diagonal. The constraints on the covariance are typically used because using a Full Covariance (FC) Model is often not warranted. Another example is the DCILT (diagonal covariances invariant to linear transformations) Model or semi-tied covariances model where the covariances are constrained to be of the form $\Sigma = A, D, A^T$, with A typically shared by several Gaussians [4, 2, 3]. This corresponds to transforming the data to an optimal feature space (in a state-dependent manner) using A and modeling using the diagonal covariance D, in this space.

3. FACTOR ANALYSIS

Often in speech recognition the DC model is used even though it is known a priori that the dimensions of the sample z_t are known to be correlated (e.g., cepstral features). Factor analysis is one approach to add more flexibility to the DC Model to capture these correlations with few parameters. The basic idea in a Factor Analyzed Covariance (FAC) Model is to estimate covariances of the form

$$\Sigma = \Lambda \Lambda^T + \Psi,$$

where $\Psi$ is a diagonal matrix and $\Lambda$ is a rectangular matrix with typically much fewer columns than rows. $\Lambda$ is referred to as the factor loading matrix and each column of $\Lambda$ is referred to as a factor. If $\Lambda$ is zero the FAC Model reduces to a DC Model. If $\Lambda$ has k factors then the FAC Model roughly tries to model the off-diagonal terms in $\Sigma$ using a rank-k matrix ($\Lambda \Lambda^T$). The FAC model also corresponds to an additive decomposition of the data into independent components - a non-squared component and a uniqueness (independent variance) component: $z_t = c + u_t$, where $c$ has zero mean and covariance $\Lambda \Lambda^T$, while $u_t$ is distributed $N(0, I)$. The hope is that with very few factors ($k \leq d$) a very good model of the covariance of the underlying Gaussian distribution is obtained. The non-squared component can be viewed as being generated by an underlying zero-mean unit covariance process. In other words,

$$z_t = \Lambda_t c_t + u_t,$$

where $z_t$ is distributed $N(0, I)$ and $u_t$ is distributed $N(\mu, \Psi)$. This viewpoint and the realization that the $z_t$'s

"specializing the FACT modeling technique to data where the constraints on the data are Markovian, we obtain a factor analysis model of the data with Gaussian observations. This description also includes HMMs with Gaussian mixture observations if the underlying state-sequence is the Gaussian mixture component sequence. Whatever be the underlying p(s), the p(x_t|s) is then completely described by state means (μ_i) and covariances (Σ_i). In practical applications the estimation of the means and covariances are constrained. Well-known examples are Linear Discriminant Analysis (where means are constrained to lie in a lower dimensional space $\text{Span}(\mu_i) = k \leq d$) and Diagonal Covariance (DC) Modeling where covariances are constrained to be diagonal viz., $\Sigma_i = D_i, D_i$ diagonal. The constraints on the covariance are typically used because using a Full Covariance (FC) Model is often not warranted. Another example is the DCILT (diagonal covariances invariant to linear transformations) Model or semi-tied covariances model where the covariances are constrained to be of the form $\Sigma = A, D, A^T$, with A typically shared by several Gaussians [4, 2, 3]. This corresponds to transforming the data to an optimal feature space (in a state-dependent manner) using A and modeling using the diagonal covariance D, in this space.

3. FACTOR ANALYSIS

Often in speech recognition the DC model is used even though it is known a priori that the dimensions of the sample z_t are known to be correlated (e.g., cepstral features). Factor analysis is one approach to add more flexibility to the DC Model to capture these correlations with few parameters. The basic idea in a Factor Analyzed Covariance (FAC) Model is to estimate covariances of the form

$$\Sigma = \Lambda \Lambda^T + \Psi,$$

where $\Psi$ is a diagonal matrix and $\Lambda$ is a rectangular matrix with typically much fewer columns than rows. $\Lambda$ is referred to as the factor loading matrix and each column of $\Lambda$ is referred to as a factor. If $\Lambda$ is zero the FAC Model reduces to a DC Model. If $\Lambda$ has k factors then the FAC Model roughly tries to model the off-diagonal terms in $\Sigma$ using a rank-k matrix ($\Lambda \Lambda^T$). The FAC model also corresponds to an additive decomposition of the data into independent components - a non-squared component and a uniqueness (independent variance) component: $z_t = c + u_t$, where $c$ has zero mean and covariance $\Lambda \Lambda^T$, while $u_t$ is distributed $N(0, I)$. The hope is that with very few factors ($k \leq d$) a very good model of the covariance of the underlying Gaussian distribution is obtained. The non-squared component can be viewed as being generated by an underlying zero-mean unit covariance process. In other words,

$$z_t = \Lambda_t c_t + u_t,$$

where $z_t$ is distributed $N(0, I)$ and $u_t$ is distributed $N(\mu, \Psi)$. This viewpoint and the realization that the $z_t$'s
can be considered latent variables leads one to an EM algorithm for FAC Model parameter estimation [1, 7, 8]. The FAC Model was recently applied for modeling the HMM states for speech recognition [9, 10].

4. FACILIT MODEL

Factor analysis is a special case of constrained ML modeling with Gaussian distributions. Wherever there are constraints it is important check if the constraints are invariant to linear transformation of the data. If the constraints are not invariant to linear transformations, then one can find an optimal linear transformation of the data so that the constraints are optimally satisfied to the extent possible in the transformed space [2]. Clearly the FAC Model is not invariant to linear transformations of the data since after a linear transformation the uniqueness component will not have a diagonal covariance. The FAC Model corresponds to optimally transforming the data (possibly in a state dependent manner) prior to modeling using a FAC Model.

In FACILIT covariances are constrained to be of the form:
\[ \Sigma = \Lambda \Lambda^\top + A^{-1} \Psi (A^{-1})^\top, \]
where \( \Lambda \) is the factor loading matrix, \( \Psi \) is the diagonal uniqueness matrix, and \( A \) is the feature transformation matrix. This corresponds to having a FAC Model of linearly transformed data. Indeed if data from each state is transformed using a matrix \( A_i \) then the FAC model corresponds to the decomposition:
\[ x_t = A_i z_t + u_{i,t}, \]
which gives rise to covariances of the form \( \Sigma_t = \Lambda_i \Lambda_i^\top + A_i^{-1} \Psi A_i^{-\top} \). Thus FACILIT is a generalization of both factor analysis (where \( A_i = I \)) and semi-tied covariance modeling (where \( \Lambda_i = 0 \)).

It turns out that in FACILIT modeling, data drawn from a set of Gaussians with covariances constrained as above can be represented in the following fashion:
\[ z_t = A_i z_t + \mu_i + \epsilon_{i,t} \]
where \( z_t \) comes from Gaussian \( z_t \) at time \( t \) and \( \epsilon_{i,t} \) is an unobserved Gaussian random variable distributed as \( N(0, I) \) and \( \mu_i \) is \( N(0, \Psi_i) \). In a FACILIT Model the conditional distribution of \( z_t \) given \( t \) and \( z_s \) is
\[ p(z_t | z_s) = \frac{\exp \left[ -\frac{1}{2} (z_t - A_i z_s - \mu_i)^\top \Psi_i^{-1} (z_t - A_i z_s - \mu_i) \right]}{\sqrt{\det (2\pi A_i^\top \Psi_i A_i)}}. \]
\[ p(x_t | z_s) = \frac{1}{\sqrt{2 \pi \sigma^2 t}} \exp \left[ -\frac{(x_t - z_s - \epsilon_t)^2}{2 \sigma^2 t} \right]. \]

Considering \( s \) and \( z_t \) as latent variables (for all \( t \)) we obtain an EM algorithm for ML estimation of FACILIT parameters viz. \( \Lambda_i, \Psi_i, \mu_i, A_i \).

If each Gaussian has its own \( \Lambda_i, \Psi_i, A_i \), then the ML estimates are trivially seen to be \( \hat{\Lambda}_i = E_i, \hat{\Psi}_i = E_i ^\top, \hat{A}_i = V_i ^\top V_i \) where \( V_i ^\top V_i \) is the eigendecomposition of the sample covariance of data from state \( i \). However, if \( A_i = I \), then there exists non-trivial ML solutions for \( \Lambda_i, \Psi_i \) - this is akin to standard factor analysis for a single Gaussian variable. More generally, if \( A_i \) is shared then there is no trivial solution. The main contribution of this paper is an EM algorithm for this form of covariance modeling allowing for the general case where \( \Lambda_i, \Psi_i, A_i \) are shared independently by arbitrary disjoint collections of Gaussians or states. For the rest of the paper we assume that \( S = \cap s \in S \) \( L_s = \cup s \in L_s \) \( T_s = \cup s \in T_s \) are independent partitions of \( S \) respectively corresponding to the sharing of \( \Lambda_i \), \( \Psi_i \) and \( A_i \).

5. THE FACILIT EM ALGORITHM

The goal is to maximize the likelihood of the data \( p(x) \), with respect to \( \mu_i, \Lambda_i, \Psi_i, A_i \) the parameters in the model. The complete data for this problem is given by the triplet \((x, z, s)\); the data, the hidden factors and the underlying state sequence. Using conditional independence of observations given hidden variables
\[ p(x, z, s) = p(s) p(z | s) p(x | z, s) = p(s) p(z | s) T_{t=1} p(x_t | z_t, s_t). \]

The Gaussian parameters depend only on \( T_{t=1} p(x_t | z_t, s_t) \), while parameters modeling the state sequence process are in \( p(s) \). The posterior distribution of the latent variables is
\[ p(x, s | x) = p(s | x) p(x | s, x). \]

EM one computes the \( Q \) function which is the expected value of the log likelihood of the complete data with respect to the posterior distribution on the hidden variables.

As for the Gaussian parameters it suffices to consider \( p(x, z, s) \) instead of \( p(x, z, s) \). If \( \theta = \{ \mu_i, \Lambda_i, \Psi_i, A_i \} \) and \( \theta = \{ \mu_i, \Lambda_i, \Psi_i, A_i \} \) are the current and new (to be estimated) values of the parameters, then from Eqn. 1,
\[ Q(\theta, \theta) = E_{\theta} \left[ \log \prod_{t=1}^{T} p(x_t | z_t, s_t) \right] \]
\[ = E_{\theta} \left[ \log \prod_{t=1}^{T} \prod_{s \in S} p(s) [p(x_t | z_t, s_t)]^{1 \{z_t \in s\}} \right] \]
\[ = \sum_{t=1}^{T} \sum_{s \in S} E_{\theta} [\log p(s)] E_{\theta} [\log p(x_t | z_t, s)] \]
\[ = \sum_{t=1}^{T} \sum_{s \in S} \gamma(t) E_{\theta} [\log p(s)] E_{\theta} [\log p(x_t | z_t, s)] \]
\[ = \sum_{t=1}^{T} \sum_{s \in S} \gamma(t) E_{\theta} [\log p(s)] E_{\theta} [\log p(x_t | z_t, s)], \]
where \( \gamma(t) = p(s_t = s | x) \) is the posterior probability of being in state \( s \) at time \( t \) given the old value of the parameters.

EM re-estimation formulæ for the parameters are obtained by setting the derivative of \( Q(\theta, \theta) \) with respect to the parameters \( \theta \) equal to zero. Solving the resulting set of simultaneous non-linear equations gives new values of the parameters. In order to express re-estimation formulæ we introduce several convenient variables. Firstly note that \( E_{\theta} [\log p(x_t | z_t, s)] \) depends on \( E_{\theta} [z_t | x, s] \) and \( E_{\theta} [z_t | x, s, z], \) which are given respectively by
\[ E_{\theta} [z_t | x, s] = \beta(s, \hat{\mu}_i), \]
\[ E_{\theta} [z_t | x, s] = \beta(s, \hat{\mu}_i). \]
and
\[ E_z[x_i,z_i,z_i] = I - \beta_s \hat{A}_s + E_z[x_i,z_i] E_z[x_i,z_i] \cdot \] (4)

where \( \beta_s = \hat{A}_s (\Psi + \hat{A}_s \hat{A}_s)^{-1} \). Now define the following statistics:
\[
\langle 1 \rangle_i = \sum_{t=1}^T \gamma_i(t). \] (5)
\[
\langle x_i \rangle = \sum_{t=1}^T \gamma_i(t) x_t. \] (6)
\[
\langle x_i z_i \rangle = \sum_{t=1}^T \gamma_i(t) x_t z_t. \] (7)
\[
\langle x_i z_i \rangle = \sum_{t=1}^T \gamma_i(t) E_z[x_i z_i], \] (8)
\[
\langle x_i z_i \rangle = \sum_{t=1}^T \gamma_i(t) E_z[x_i z_i]. \] (9)
\[
\langle x_i z_i \rangle = \sum_{t=1}^T \gamma_i(t) \Psi^{-1} A_i \mu_i(z_i). \] (10)
\[
\Psi^{-1} A_i z_i = \sum_{t=1}^T \gamma_i(t) \Psi^{-1} A_i E_z[z_i]. \] (11)
\[
\langle A_i z_i \rangle = \sum_{t=1}^T \gamma_i(t) A_i z_t. \] (12)
\[
\langle A_i z_i \rangle = \sum_{t=1}^T \gamma_i(t) A_i z_t. \] (13)
\[
\langle A_i z_i \rangle = \sum_{t=1}^T \gamma_i(t) A_i z_t. \] (14)
\[
V_{ar}[A_i z_i] = \langle A_i z_i \rangle - \frac{\langle A_i z_i \rangle}{\langle 1 \rangle_i} \frac{\langle A_i z_i \rangle}{\langle 1 \rangle_i}. \] (15)
\[
Cov(z_i) = \langle z_i z_i \rangle - \langle z_i \rangle \langle z_i \rangle. \] (16)
\[
Cov(z_i) = \langle z_i z_i \rangle - \langle z_i \rangle \langle z_i \rangle. \] (17)
\[
Cov(z_i) = \langle z_i z_i \rangle - \langle z_i \rangle \langle z_i \rangle. \] (18)
\[
G^{(e)} = \sum_{i \in A_s} \frac{1}{\Psi_s(i,i)} \left[ (1) |Cov(z_i) + \Lambda_i Cov(z_i)A_i \right] \right]. \] (19)

The re-estimation formulae are as follows:

\textbf{A estimation - Case 2} If \( A_i \) are not shared among states or are shared at a "lower level" than either the \( A' \)’s or \( \Psi \)’s i.e., if for all \( s \in L_i, A_i = A_s \), for some \( r \) and \( \Psi_s = \Psi_p \) for some \( p \) then
\[
A_i = \left( \sum_{i \in C_i} \langle x_i z_i \rangle - \mu_i \langle z_i \rangle \right) \left( \sum_{i \in C_i} \langle x_i z_i \rangle \right)^{-1}. \] (20)

\textbf{Optimal Feature Space \( (A_s) \) estimation} The rows of \( A_s \) are obtained in an iterative fashion using the following formula:
\[
a_i = c_i G_s^{(e)} - \frac{1}{c_i} \sum_{i \in A_s} \langle 1 \rangle_i. \] (23)

where \( c_i \) are the cofactors of the \( i \)th row of \( A_s \).

\textbf{5.1. Sufficient Statistics}

Since we are estimating Gaussians the sufficient statistics for the estimation are the zeroth, first and second order statistics of the data for each Gaussian. In other words it suffices to have the statistics in Eqn. 6-Eqn. 7. If these statistics are available, then FAC Model can be readily obtained using EM. In some problems where storage of the second order statistics is prohibitive (e.g., speech recognition), one can compute the alternate statistics Eqn. 8-Eqn. 10. However, with these statistics only one iteration of EM can be performed since the \( z \) statistics change with each iteration. Therefore, repeated alignments of the training data is required in this mode of computation. There is an evident tradeoff between computational and storage requirements.

\textbf{5.2. Likelihood Computation}

The likelihood computation for FAC Modeling is computationally intensive as PC Modeling since the inverse covariance (which is full) is required in both cases. However, in a FAC Model the storage requirements can be reduced by using the Sherman-Morrison-Woodbury formula for inversion of rank-one updates of a matrix (recall the FAC covariance is a rank \( k \) update of \( \Psi \)). Indeed for a FAC Model we have
\[
\Sigma^{-1} = \Psi^{-1} - \Psi^{-1} A_s (I + A_s \Psi^{-1} A_s)^{-1} \Psi^{-1} A_s. \] (24)

FACILIT model likelihood computation is identical but for the fact that the data has to be transformed in a class-dependent fashion prior to likelihood computation. Specifically, in the global HLT case (where there is a single global transformation) the data can be transformed a priori and a standard FAC Model used.
All these systems were obtained by boot-strapping from a number of factors - again perhaps indicating over-training. The best results are obtained for the two-factor case with the least number of parameters - i.e., with maximal sharing. With sharing fixed at the phone level the number of factors was changed to 3, 4, and 6 and the results seem to degrade with number of factors - again perhaps indicating over-training. All these systems were obtained by boot-strapping from a diagonal covariance system.

### 5.3. Initialization of Parameters

The EM algorithm leads to a local maximum of likelihood and as such the solution depends on the initial value of the parameters. If we start with a DC Model a good starting point - one might suppose - is to initialize the \( \mu_s \)'s and \( \Sigma_s \)'s with the means and diagonal covariances of the DC Model and to initialize the \( A_i \)'s to zero. Unfortunately, this is a local minimum of the likelihood function and hence a stationary point. Therefore, typically \( A_i \) are randomly initialized with small values. One possibility is to use Joreskog's method as suggested in [6] for standard factor analysis.

### 6. EXPERIMENTAL RESULTS

The first set of experiments were run on an internal test database of about 1200 words each from 10 speakers. The training data for this database is from an internal database of read speech and the acoustic models used 52 phones, 156 sub-phonetic units, about 27K HMM states and about 32K Gaussians. The goal was to study the effect of sharing factors at various levels of tying: phone-level, sub-phonetic unit (arc) level, and context-dependent sub-phonetic unit (or HMM state) level. The results shown in Table 6 indicate that there may be an over-training problem. The best results are obtained for the two-factor case with the least number of parameters - i.e., with maximal sharing. With sharing fixed at the phone level the number of factors were changed to 3, 4, and 6 and the results seem to degrade with number of factors - again perhaps indicating over-training. All these systems were obtained by boot-strapping from a diagonal covariance system.

To study the effectiveness of FAC Modeling on spontaneous speech data a series of experiments were conducted on a Voicemail Test Data. The training data consists of about 20 hours of Voicemail data and the test data consists of about 45 Voicemail messages each about 20 seconds duration. The acoustic models in these experiments used an augmented phone set (two additional phones added to model clicks and disfluencies), 162 sub-phonetic units, 3K HMM states and 73K Gaussians. The results are shown in Table 6.

Currently experiments are underway to study the effects of using \( A_i \)'s shared at the global, phone and sub-phonetic unit levels. Also, it is fairly straightforward to decide on the number of factors in a data-driven fashion based on the relative gain in likelihood.

---

### 7. CONCLUSION

This paper introduces FACILT - a new covariance modeling technique for Gaussian distributions - and presents an EM algorithm for estimating the model parameters in FACILT. FACILT is factor analysis in optimal feature spaces and hence is a generalization of classical factor analysis. Semi-tied covariance modeling and diagonal covariance modeling can be seen as a special cases of FACILT.

### 8. REFERENCES


---

### Table 1: FAC Model Word Error Rate on Read Speech Test Data: 32K Gaussians, 2.7K HMM states

<table>
<thead>
<tr>
<th>Sharing</th>
<th>k=0</th>
<th>k=2</th>
<th>k=4</th>
<th>k=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>12.0</td>
<td>12.1</td>
<td>12.3</td>
<td>12.3</td>
</tr>
<tr>
<td>Arc</td>
<td>12.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMM State</td>
<td>12.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>12.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: FAC Model Word Error Rate on Spontaneous Speech Test Data: 70K Gaussians, 3K HMM states

<table>
<thead>
<tr>
<th>Sharing</th>
<th>k=0</th>
<th>k=2</th>
<th>k=4</th>
<th>k=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>39.2</td>
<td>39.4</td>
<td>40.0</td>
<td></td>
</tr>
<tr>
<td>HMM State</td>
<td>39.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>48.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>