FACTORED SPARSE INVERSE COVARIANCE MATRICES

Jeff A. Bilmes
<bilmes@ee.washington.edu>
Dept. of Electrical Engineering, University of Washington
418 EE/CS Bldg, Box 352500, Seattle, WA 98195, USA

ABSTRACT

Most HMM-based speech recognition systems use Gaussian mixtures as observation probability density functions. An important goal in all such systems is to improve parsimony. One method is to adjust the type of covariance matrices used. In this work, factored sparse inverse covariance matrices are introduced. Based on $U'DU$ factorization, the inverse covariance matrix can be represented using linear regressive coefficients which 1) correspond to sparse patterns in the inverse covariance matrix (and therefore represent conditional independence properties of the Gaussian), and 2), result in a method of partial tying of the covariance matrices without requiring non-linear EM update equations. Results show that the performance of full-covariance Gaussians can be matched by factored sparse inverse covariance Gaussians having significantly fewer parameters.

1. INTRODUCTION

Most state-of-the-art speech recognition systems represent the joint distribution of features for each utterance using hidden Markov models (HMMs) with multivariate Gaussian mixture observation densities [15]. An important goal for designers of automatic speech recognition (ASR) systems is to achieve a high level of performance while minimizing the number of parameters used by the system. One way of controlling the number of parameters is to adjust the structure of the covariance matrix used by each Gaussian mixture component. Traditionally, the choice is made between either diagonal or full covariance matrices.

With diagonal covariance matrices, all off-diagonal matrix elements are set to zero. For a single Gaussian component, this means the random variables are assumed to be statistically independent. With mixtures of diagonal-covariance Gaussians, dependencies between random variables can be represented, but complex distributions potentially require a large number of components. The alternative, requiring many more parameters, has been to use full covariance matrices where each component corresponds to a more complex distribution. It has been demonstrated [10] that, at least for the standard features used for speech recognition (cepstral features), representing correlation explicitly by including non-zero off-diagonal covariance elements can improve word accuracy over a simple mixture of diagonal-covariance Gaussians. To avoid the complexity of more parameters, this suggests a compromise should be used between diagonal and full covariance matrices.

There are a variety of choices for covariance structure other than diagonal or full, some of which have been previously used as HMM state-conditioned observation densities. Two examples include block-diagonal [8] and banded-diagonal matrices. Another method often used by ASR systems to reduce parameters (and thereby increase estimation robustness) is tying, where certain parameters are shared amongst a number of different models. Accordingly, various matrix decomposition methods of the form $C = A'DA$ (where D is diagonal and A is an arbitrary matrix) have been applied to covariance matrices along with different styles of partial parameter tying [5, 10, 16]. These methods could collectively be called partially tied covariance matrices since only a portion of the covariance matrix is not tied and remains uniquely associated with each mixture component of each HMM state.

While many statistical systems allow for a regular structure, it is becoming apparent that the use of a sparse structure is one method to eliminate the unnecessary parameters in a system. One way of controlling sparseness (and number of parameters) is by adjusting the inherent statistical dependencies made by a probabilistic model. Ideally, only the important statistical dependencies in the training data should be represented [1] and the direct relationships between the remaining random variables should be left unspecified.

Covariance matrices are no exception to this rule. In general, the location of any zeros in the inverse covariance matrix of a Gaussian distribution correspond to the conditional independence properties of that distribution. By forcing certain elements of the inverse covariance matrix to be zero, the number of parameters in the system can be reduced. This is the idea behind *covariance selection*, originally advocated in [4], described in [6, 9, 13], and proposed for speech in [1, 3]. Also, in [14], a procedure was given for learning the structure of mixtures of Bayesian networks which, in that work, corresponded to mixtures of Gaussians with sparse inverse covariance matrices.

This paper introduces two new procedures: 1) a method for covariance selection based on choosing statistical dependencies according to conditional mutual information computed using training data, and 2) a method for partially tying covariance matrices based on $U'DU$ factorization, where U is a unit upper triangular matrix and D is diagonal. This leads to an easy way of specifying the sparse patterns of the inverse covariance matrices, and also results in efficient linear EM update equations even when the covariance matrices are partially tied.

Section 2 reviews an interesting property of jointly Gaussian random variables showing that each variable is a Gaussian autoregressive process with dependencies on the other variables. It also informally shows how conditional independence properties of a Gaussian distribution are directly related to the inverse covariance matrix. Section 3 shows how, via $U'DU$ decomposition, the inverse covariance matrix can be represented as factors of linear-regressive and variance components, and how the resulting factorization can lead to a simple EM parameter optimization strategy even when the matrices are partially tied. Section 4 outlines the method used to choose the sparse inverse covariance matrix pattern. Section 5 shows how the number of parameters can be significantly reduced without greatly affecting WER performance relative to a full-covariance matrix system. Finally, section 6 concludes and discusses future work.

2. NORMAL DISTRIBUTIONS AND CONDITIONAL INDEPENDENCE

A multivariate Gaussian distribution $f(x)$ is defined as follows:

$$f(x) = |2\pi\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

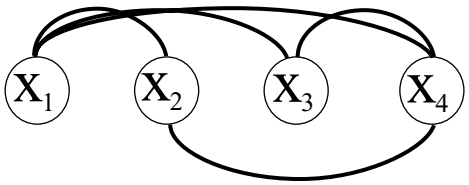

Figure 1: Zeros in the inverse covariance matrix correspond to missing edges in the graphical model, and therefore control the conditional independence properties of the Gaussian distribution. In the figure, there is no edge between X_2 and X_3 implying that $X_2 \perp\!\!\!\perp X_3 | \{X_1, X_4\}$ and also implying that $K_{23} = K_{32} = 0$.

where x is d-dimensional multivariate-vector, μ is the mean, and Σ is the covariance matrix. Note that if an element of the covariance matrix is zero (i.e., $\Sigma_{i,j} = 0$), then the two corresponding variables are marginally independent (i.e., $f(x_i, x_j) = f(x_i)f(x_j)$).

One important property of jointly Gaussian random variables is that linear transformations of and conditioning on any set of the component scalar variables preserves the Gaussian property. The chain rule of probability states that:

$$f(x) = \prod_i f(x_i | x_{1:i-1})$$

Because the factors $f(x_i|x_{1:i-1})$ are each Gaussian, any Gaussian distribution can be represented as a product of conditional Gaussian distributions. A well-known result from multivariate statistics is that $f(x_i|x_{1:i-1})$ has distribution with conditional mean:

$$\mu_{i|1:i-1} = \mu_i + \Sigma_{i,1:i-1}(\Sigma_{1:i-1,1:i-1})^{-1}(x_{1:i-1} - \mu_{1:i-1})$$

and covariance:

$$\Sigma_{i|1:i-1} = \Sigma_{ii} - \Sigma_{i,1:i-1}(\Sigma_{1:i-1,1:i-1})^{-1}\Sigma_{1:i-1,i}$$

where the notation $A_{i:j,l:m}$ refers to a partition of a matrix A containing rows i through j and columns l through m.[1] Note that x_i's dependency on $x_{1:i-1}$ is only through x_i's conditional mean – the covariance stays fixed as $x_{1:i-1}$ changes.

The conditionally Gaussian distribution of x_i given $x_{1:i-1}$ can be seen as linear regression [9, 13] since in this case:

$$x_i = \mu_i + K_{ii}^{-1}K_{i,1:i-1}(x_{1:i-1} - \mu_{1:i-1}) + \epsilon_i$$

where

$$\epsilon_i \sim \mathcal{N}(0, \Sigma_{ii} - K_{ii}^{-1}K_{i,1:i-1}\Sigma_{1:i-1,i})$$

and where $K = (\Sigma_{1:i,1:i})^{-1}$ is the inverse covariance matrix.

In this form, it can be seen that conditional independence properties of the distribution are determined by the location of zeros in the inverse covariance matrix. This is because when $K_{ij} = 0$, x_i no longer depends on x_j given the remaining variables.[2] A Gaussian distribution can also be viewed as a graphical model, where nodes in the graph correspond to the scalar random variables, and edges in the graph exist for each non-zero off-diagonal entry in the inverse covariance matrix (see Figure 1).

[1] Also, A_{ij} will refer to the element of a matrix A in row i and column j, and $A_{i,l:m}$ to columns l through m of row i.

[2] In [9], it is formally proven that $K_{ij} = 0$ if and only if $x_i \perp\!\!\!\perp x_j | x_{\{1:d\}\setminus\{i,j\}}$ where $K = \Sigma^{-1}$

3. FACTORED SPARSE INVERSE COVARIANCE MATRICES

Another way of specifying the auto-regression coefficients of a Gaussian is to decompose the inverse covariance matrix and then manipulate the outer factors so that they lie within the linear components of the exponential's Mahalanobis distance argument. This decomposition also leads to a method where the sparse patterns in the inverse covariance matrix are specified by setting regression coefficients to zero. It also allows the partial tying of the covariance matrix without needing a non-linear optimization procedure within each EM iteration.

Any positive definite matrix A has a unique decomposition into factors $U'DU$, where D is a positive diagonal matrix and U is a unit upper-triangular matrix. A unit triangular matrix is a triangular matrix that has ones along the diagonal. This implies that U is non-singular and that $det(A) = det(U'DU) = det(D)$ since $det(U) = 1$. A Gaussian density can therefore be represented as:

$$f(x) = |2\pi D|^{-1/2} e^{-\frac{1}{2}(x-\mu)'U'DU(x-\mu)}$$

where $\Sigma^{-1} = U'DU$.

If U was a general matrix, the EM update equations for such a model would require a non-linear optimization procedure to be performed within each EM iteration (see [5, 16]). The unit triangular matrices can be "brought" inside the linear terms, however, as follows:

$$\begin{aligned}
&(x-\mu)'U'DU(x-\mu) \\
&\quad = (U(x-\mu))'D(U(x-\mu)) \\
&\quad = (Ux - \tilde{\mu})'D(Ux - \tilde{\mu}) \\
&\quad = ((I-B)x - \tilde{\mu})'D((I-B)x - \tilde{\mu}) \\
&\quad = (x - Bx - \tilde{\mu})'D(x - Bx - \tilde{\mu})
\end{aligned}$$

where $U = I - B$, I is the identity matrix, B is an upper triangular matrix with zeros along the diagonal, and $\tilde{\mu} = U\mu$ is the new mean. Note that if $B_{ij} = 0$ for $j > i$, then $K_{ij} = K_{ji} = 0$ where $K = \Sigma^{-1}$.

This process transforms Σ^{-1} into a linear regression on x, without effecting the Gaussian normalization coefficient. Therefore, a full-covariance Gaussian distribution is like a conditional Gaussian distribution with conditioning variables coming from the same feature vector rather than from somewhere else in time. The same optimization procedures as those used in [1, 2, 12] can therefore be used here.

As in other matrix decompositions [5, 16], the covariance matrix for a particular mixture component and HMM state can be represented as $\Sigma^{(m)} = U^{(r)'}D^{(m)}U^{(r)}$ where $U^{(r)}$ may be tied together over a number of different components, with $D^{(m)}$ remaining uniquely associated to a particular component of an HMM state. With this decomposition, a partially tied Gaussian mixture HMM system can be trained using an EM optimization strategy that does not require an iterative non-linear optimization procedure within each EM iteration [5, 16]. This is essentially a consequence of the following formulas [1] which are used to produce the EM update equations for these models. In the first case,

$$\begin{aligned}
&\frac{\partial}{\partial B}\Big(\log(|(I-B)'D(I-B)|) + \\
&\qquad (x - Bx - \tilde{\mu})'D(x - Bx - \tilde{\mu})\Big) \\
&\qquad = -2D(x - Bx - \tilde{\mu})x'
\end{aligned}$$

whereas in the general case, the derivative

$$\frac{\partial}{\partial A}\Big(\log(|A'DA|) + (x-\mu)'A'DA(x-\mu)\Big)$$

for an arbitrary matrix A is not as easy to represent.

4. COVARIANCE SELECTION STRATEGY

With the decomposition $\Sigma^{-1} = U'DU$, choosing the locations of zeros in the inverse covariance matrix is equivalent to choosing the zeros in the triangular matrix B. This is identical to the problem encountered in [2], where the individual dependencies between feature vectors needed to be chosen. In this paper a simple strategy is proposed to choose the zeros of B: $B_{i,j}$ may be non-zero only if the conditional mutual information $I(X_i; X_j|Q = q)$ is large enough.

There are a several ways to obtain $I(X_i; X_j|Q = q)$. One strategy starts with a diagonal-covariance Gaussian HMM system, computes Viterbi paths, computes conditional mutual information (as described in [1]), sets certain elements of B to be non-zero, and trains the result. This is somewhat analogous to a forward approach [4] to covariance selection (and also similar to the forward feature selection procedure of [7]). This strategy has the advantage that it requires only a simple boot system containing relatively few but robustly estimated parameters. Adding additional dependencies can be seen as correcting deficiencies in the model as measured in according to the training data.

A problem with the above approach is that, starting from a simple system, the Viterbi paths might not be very accurate which reduces the quality of the conditional mutual information estimation. An alternative strategy then is to start with a full-covariance HMM system and perform the remaining steps in the same way. This is analogous to a backward [4] (or backward selection [7]) procedure. It has the advantage that more precise estimates of $I(X_i; X_j|Q = q)$ can be obtained resulting in better sparse matrices. The disadvantage is that it requires a trained full-covariance HMM system, the parameters of which might not be as robustly estimated as the diagonal-covariance system. There is obviously a trade-off between the two approaches. In this work, however, only the later approach is evaluated.

5. RESULTS

Speech recognition results were obtained using NYNEX PHONEBOOK, a large-vocabulary, phonetically-rich, isolated-word, telephone-speech database[11]. Data is represented using 12 MFCCs plus c_0 and deltas resulting in a $d = 26$ element feature vector sampled every 10ms. The training and test sets are as defined in [2]. Test words do not occur in the training vocabulary, so test word models are constructed using phone models learned during training. Strictly left-to-right transition matrices were used except for an optional beginning and ending silence model.

An HMM baseline system, bootstrapped using uniform segmental k-means, was developed using 42 phone models (41 monophones + silence) and three HMM states per mono-phone. Each phone model uses a mixture of 5 full-covariance Gaussians, specified by allowing non-zero values for all off-diagonal elements of the B matrices. The dictionary included with PHONEBOOK was used for all pronunciations. All training was performed using standard EM for maximum-likelihood parameter estimation.

Given the trained full-covariance system, Viterbi paths were computed for the entire training set, and conditional mutual information was calculated for each HMM state (again see [1]). The structure of the B matrices for each component was chosen by selecting a certain percentage of the possible non-zero entries of B. The structure of the B matrices for each component of an HMM state was set to be the same. Three different methods to choose a certain percentage of possible non-zero entries of B were compared. The first method chooses the entries corresponding to the greatest values of the conditional mutual information, the second chooses the entries corresponding to the smallest values, and the third chooses random entries. In all cases, the values used were 10%, 20%, ..., 90% of the possible B entries.

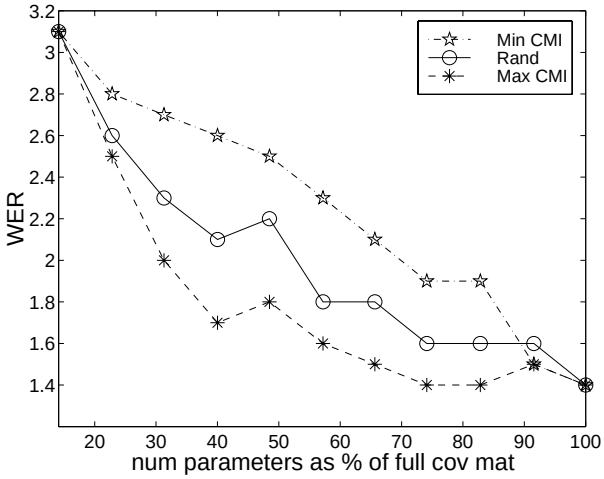

Figure 2: 75 Word lexicon size.

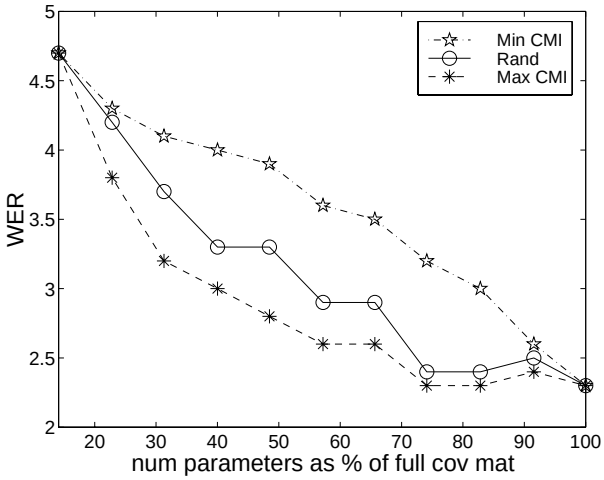

Figure 3: 150 Word lexicon size.

Results are plotted in Figures 2 through 5 which show the word error (WER) performance on test sets with different lexicon sizes (the perplexity in this case is equal to the lexicon size). The results are plotted against the percentage of parameters relative to the full-covariance result. Each point in the plot corresponds to the WER evaluation of a system obtained by choosing non-zero entries for the B matrices of each component of each HMM state, and training and then testing the result. The left-most point on the plots corresponds to a system with diagonal covariance matrices. Such a system requires only 14% of the parameters of a full covariance matrix. The right most point on the plot corresponds to full covariance matrices (i.e., B matrices), and has an x-axis value of 100%.

The plots clearly show that in all cases adding entries to the diagonal inverse covariance matrix decreases WER. When they are added according to the maximum conditional mutual information procedure, the rate of WER decrease is greatest, followed by ran-

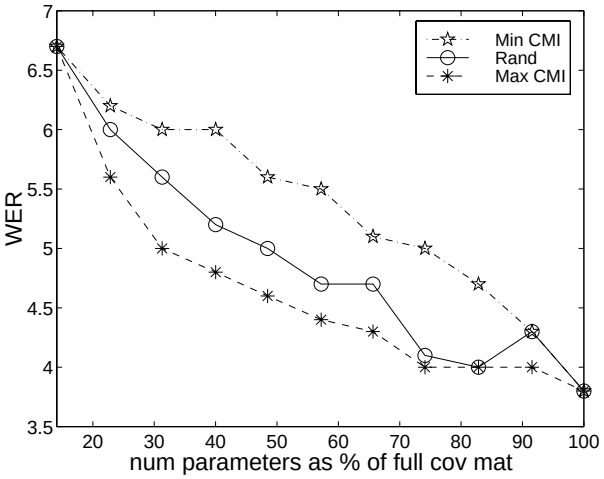

Figure 4: 300 Word lexicon size.

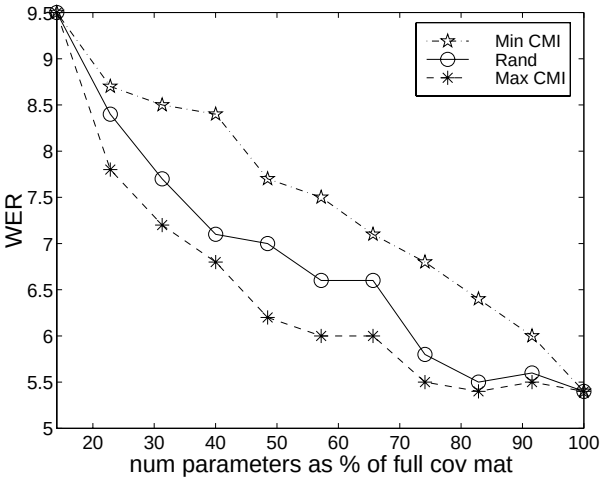

Figure 5: 600 Word lexicon size.

dom and then by minimum conditional mutual information. This is as expected: choosing entries according to mutual information chooses first the more important dependencies represented by the Gaussian. It is interesting to note that the random procedure does reasonably well. The minimum procedure performs the worst, as it first chooses dependencies that are least important. In general, the plots show that only about 70% of the parameters of a full covariance system are needed to achieve the same performance.

6. DISCUSSION

In the above experiments, the conditional mutual information was computed at the HMM state level and was used to adjust the inverse covariance matrix's sparse pattern for all of the state's mixture components. Permitting B_{ij} to be non-zero will certainly allow a dependence between x_i and x_j, but this dependence could be superfluous since, via the Gaussian mixture, the capability to represent a dependence between x_i and x_j might already implicitly exist. An alternative procedure could compute the quantity $I(X_i; X_j|Q = q, M = m)$ where m is a mixture component, and the B matrix for each mixture component is adjusted individually. Mutual information in this case would tell us more about the dependence than can be represented by the linear dependence in each Gaussian. Two alternatives are to use a poorer measure of dependence (e.g., correlation), or to use a richer model at each component. For the former case, one could just set the elements of each B in the full-covariance system to zero that fall below a threshold, although this was not attempted in this paper. The later case will be investigated in future work.

Additional future work will use discriminative mutual information [2] to select covariance structure. This will perhaps provide models that perform as well with even fewer parameters. Also, the partially tied covariance matrices introduced in this paper were not yet tested.

The author would like to acknowledge the International Computer Science Institute at which the computational experiments were performed. Many thanks to Les Atlas, Becky Bates, and Katrin Kirchhoff who reviewed drafts of this paper.

REFERENCES

[1] J. Bilmes. *Natural Statistic Models for Automatic Speech Recognition*. PhD thesis, U.C. Berkeley, Dept. of EECS, CS Division, 1999.

[2] J.A. Bilmes. Buried Markov models for speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, March 1999.

[3] S.S. Chen and R.A. Gopinath. Model selection in acoustic modeling. *EUROSPEECH*, 1999.

[4] A.P. Dempster. Covariance selection. *Biometrics*, 28:157–75, March 1972.

[5] M.J.F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Trans. Speech and Audio Proc.*, 7(3), 1999.

[6] M. Knuiman. Covariance selection. *Suppl. Advances in Applied Probability*, 10:123–130, 1978.

[7] D. Koller and M. Sahami. Toward optimal feature selection. In *Machine Learning: Proc. of the 13th International Conference*. Morgan Kaufmann, 1996.

[8] R. Koshiba, M. Tachimori, and H. Kanazawa. A flexible method of creating HMM using block-diagonalization of covariance matrices. *Int. Conf. on Spoken Language Proc.*, 1998.

[9] S.L. Lauritzen. *Graphical Models*. Oxford Science Publications, 1996.

[10] A. Ljolje. The importance of cepstral parameter correlations in speech recognition. *Computer Speech and Language*, 8:223–232, 1994.

[11] J. Pitrelli, C. Fong, S.H. Wong, J.R. Spitz, and H.C. Lueng. PhoneBook: A phonetically-rich isolated-word telephone-speech database. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1995.

[12] A.B. Poritz. Linear predictive hidden Markov models and the speech signal. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 1291–1294, 1982.

[13] R.E. Roger. Sparse inverse covariance matrices and efficient maximum likelihood classification of hyperspectral data. *Int. J. of Remote Sensing*, 17(3):589–613, 1996.

[14] B. Thiesson, C. Meek, D. Chickering, and D. Heckerman. Learning mixtures of Bayesian networks. Technical Report MSR-TR-97-30, Microsoft Research, 1998.

[15] S. Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine*, 13(5), 1996.

[16] K.-H. Yuo and H.-C. Wang. Joint estimation of feature transformation parameters and gaussian mixture model for speaker identification. *Speech Communication*, 3(1):211–226, July 1999.