LOW RESOURCE SPEECH RECOGNITION WITH AUTOMATICALLY LEARNED SPARSE INVERSE COVARIANCE MATRICES

Weibin ZHANG, Pascale FUNG

HKUST
Human Language Technology Center
Department of Electronic and Computer Engineering
University of Science and Technology, Clear Water Bay, Hong Kong
wbzhang@ust.hk, pascale@ece.ust.hk

ABSTRACT

Full covariance acoustic models trained with limited training data generalize poorly to unseen test data due to a large number of free parameters. We propose to use sparse inverse covariance matrices to address this problem. Previous sparse inverse covariance methods never outperformed full covariance methods. We propose a method to automatically drive the structure of inverse covariance matrices to sparse during training. We use a new objective function by adding L1 regularization to the traditional objective function for maximum likelihood estimation.The graphic lasso method for the estimation of a sparse inverse covariance matrix is incorporated into the Expectation Maximization algorithm to learn parameters of HMM using the new objective function. Experimental results show that we only need about 25% of the parameters of the inverse covariance matrices to be nonzero in order to achieve the same performance of a full covariance system. Our proposed system using sparse inverse covariance Gaussians also significantly outperforms a system using full covariance Gaussians trained on limited data.

Index Terms— sparse inverse covariance matrix, speech recognition, graphic lasso,expectation maximization

1. INTRODUCTION

Acoustic models trained tend to over fit when there is not enough training data as the model has too many parameters relative to the amount of observed data, especially when full covariance matrices are used. An over-fitting model will generalize poorly to unseen test data. Typical approaches to solve this problem include using state tying methods where data from both resource poor and resource rich genre of speech are shared.

Another approach is to explicitly compact the models. Various heuristic methods such as subspace Gaussian mixture models (SGMM [1]) and subspace for precision and mean models (SPAM [2]) were previously proposed for model compaction. [3] first proposed to model the inverse covariance matrices with a sparsity structure. If the ijth component of the inverse covariance matrix is zero, then variables i and j are conditionally independent, given the other variables. Sparse inverse covariance matrices also lead to computational advantage since the Gaussian likelihoods are evaluated as a quadratic form determined by the inverse covariance matrices. Unfortunately no results related to sparse inverse covariance matrices were presented in [3]. [4] proposed to heuristically choose the locations of zeros in the inverse covariance matrices based on the conditional mutual information of two random variables. Results showed that only about 70% of the parameters of a full covariance system are needed to achieve the same performance. However, in [4] the sparse inverse covariance systems never outperform the full covariance systems if the locations of zeros in the inverse covariance matrices are set before training, using the methods proposed.

In machine learning, regularization terms are usually added to the objective function to penalize complex models or to impose prior knowledge. One popular type of regularization is l_1 regularization which results in sparse models. These simple sparse models tend to generalize well to unseen test data. Recently, [5] solved the problem of estimating a sparse inverse matrix when the training data is assumed to be drawn from a Gaussian distribution and used it for model selection. [6] proposed a more efficient way– the *graphic lasso*– for the estimation of a sparse inverse covariance matrix. By using *graphic lasso* and Expectation Maximization (EM [7]), we propose a method to automatically learn the sparse structure of the inverse covariance matrices for low resource acoustic model training in this paper.

The rest of this paper is organized as follows. In section 2, we show how the sparse inverse covariance matrices can be learned automatically. We elaborate the new objective function and how to maximize it using EM algorithm. The experiment results are given in section 3. Finally the conclusion and future work are stated in section 4.

2. AUTOMATIC LEARNING OF SPARSE INVERSE COVARIANCE MATRICES

We propose to automatically learn sparse inverse covariance matrices using Expectation Maximization and *graphic lasso* as follows. An l_1 penalty term is added to the traditional objective function (see equation(1)) for maximum likelihood estimation (MLE) in order to automatically drive the structure of the inverse covariance matrices to sparse. This new objective function is then maximized using EM algorithm. Details are given below.

Suppose we have a HMM-based acoustic model $P(\mathbf{O}|\mathbf{\Theta})$ that is governed by a set of parameters $\mathbf{\Theta}$. We also have observation sequences $\mathbf{O}$ that is supposed to be drawn from this distribution. The traditional maximum likelihood estimation method is to find the parameters that maximize the log likelihood function $\log(P(\mathbf{O}|\mathbf{\Theta}))$, that is

$$\hat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}}{\operatorname{argmax}}\{\log(P(\mathbf{O}|\mathbf{\Theta}))\}. \tag{1}$$

We propose to add an l_1 penalty in order to automatically drive the learned precision matrices (inverse of the covariance matrices) to sparse. Thus the following new objective function to be maximized is considered

$$\mathcal{L}(\mathbf{\Theta}) = \log(P(\mathbf{O}|\mathbf{\Theta})) - \sum_{s=1}^{S}\sum_{l=1}^{M_s}\lambda_{sl}||\mathbf{C}_{sl}||_1, \tag{2}$$

where $||\cdot||_1$ denotes the l_1 norm of a matrix (i.e. the sum of the absolute values of the elements of the matrix), S is the number of states in HMM, M_s is the number of Gaussian components in state s, $\mathbf{C}_{sl}$ is the precision matrix of the l^{th} mixture component in state s and λ_{sl} is a user defined hyperparameter. It can be seen that maximizing the new objective function is equivalent to a *maximum a posteriori* (MAP) procedure by using the following Laplace priors on the inverse covariance matrices

$$p(\mathbf{C}_{sl}|\lambda_{sl}) = \frac{\lambda_{sl}}{2}\exp(-\lambda_{sl}||\mathbf{C}_{sl}||_1).$$

Since the new objective function (2) is expensive to evaluate, exact maximization is intractable. We try to use the EM algorithm to find a local optima [8]. The procedure of using EM is first to obtain an auxiliary function that is the lower bound of the objective function. Iteratively maximizing (or increasing) this auxiliary function guarantees the increase of the target objective function. The following auxiliary function is defined

$$\begin{aligned} Q(\mathbf{\Theta}, \mathbf{\Theta}') = &\sum_{q\in\mathfrak{Q}}\sum_{m\in\mathfrak{M}} logP(O,q,m|\mathbf{\Theta})P(O,q,m|\mathbf{\Theta}') \\ &- \sum_{s=1}^{S}\sum_{l=1}^{M_s}\lambda_{sl}||\mathbf{C}_{sl}||_1, \end{aligned} \tag{3}$$

where $P(O,q,m|\mathbf{\Theta})$ is the likelihood of generating $\mathbf{O}$ using the state sequence q and the Gaussian components in each state indicated by m; $\mathbf{\Theta}'$ is our previous estimate of the parameters. By using Jensen's inequality, we can get

$$\mathcal{L}(\mathbf{\Theta}) - \mathcal{L}(\mathbf{\Theta}') \geq Q(\mathbf{\Theta}, \mathbf{\Theta}') - Q(\mathbf{\Theta}', \mathbf{\Theta}'), \tag{4}$$

meaning that $Q(\mathbf{\Theta}, \mathbf{\Theta}') - Q(\mathbf{\Theta}', \mathbf{\Theta}')$ is a lower bound of $\mathcal{L}(\mathbf{\Theta}) - \mathcal{L}(\mathbf{\Theta}')$. Thus if a value $\mathbf{\Theta}$ satisfies $Q(\mathbf{\Theta}, \mathbf{\Theta}') > Q(\mathbf{\Theta}', \mathbf{\Theta}')$,then $\mathcal{L}(\mathbf{\Theta}) > \mathcal{L}(\mathbf{\Theta}')$. Therefore the mode of the objective function $\mathcal{L}(\mathbf{\Theta})$ can be estimated by iteratively maximizing the auxiliary function $Q(\mathbf{\Theta}, \mathbf{\Theta}')$.

Expanding $logP(O,q,m|\mathbf{\Theta})P(O,q,m|\mathbf{\Theta}')$ in equation (3) as that in [8] and group different types of parameters (e.g. initial distributions and transition probabilities) together, we can see that the training procedures for the HMM parameters are the same as MLE method except the precision matrices. Further more, by eliminating irrelevant constants, the auxiliary function for the estimation of precision matrices is

$$\begin{aligned} Q(\mathbf{C}, \mathbf{C}') &= \sum_{s=1}^{S}\sum_{l=1}^{M_s}\sum_{t=1}^{T}(\log|\mathbf{C}_{sl}| - tr((o_t - \mu_{sl})\cdot \\ &\qquad (o_t - \mu_{sl})'\mathbf{C}_{sl}))\gamma_{slt} - \sum_{s=1}^{S}\sum_{l=1}^{M_s}\lambda_{sl}||\mathbf{C}_{sl}||_1 \\ &= \sum_{s=1}^{S}\sum_{l=1}^{M_s}(\gamma_{sl}\log|\mathbf{C}_{sl}| - \gamma_{sl}tr(\mathbf{S}_{sl}\mathbf{C}_{sl}) - \lambda_{sl}||\mathbf{C}_{sl}||_1) \\ &= \sum_{s=1}^{S}\sum_{l=1}^{M_s}\gamma_{sl}(\log|\mathbf{C}_{sl}| - tr(\mathbf{S}_{sl}\mathbf{C}_{sl}) - \frac{\lambda_{sl}}{\gamma_{sl}}||\mathbf{C}_{sl}||_1) \\ &= \sum_{s=1}^{S}\sum_{l=1}^{M_s}\gamma_{sl}Q(\mathbf{C}_{sl}, \mathbf{C}'_{sl}) \end{aligned} \tag{5}$$

where $|\cdot|$ denotes the determinant of a matrix, $tr(\cdot)$ denotes the trace of a matrix and the quantity γ_{slt}

$$\gamma_{slt} = p(q_t = s, m_{st} = l|O, \mathbf{\Theta}')$$

is the posteriori probability of occupying the l^{th} Gaussian component in state s at time t given observation $\mathbf{O}$ is generated; γ_{sl} is defined as

$$\gamma_{sl} = \sum_{t=1}^{T}\gamma_{slt}.$$

μ_{sl} and C_{sl} are the mean and precision matrix of the l^{th} Gaussian component in state s. $\mathbf{S}_{sl}$ and $Q(\mathbf{C}_{sl}, \mathbf{C}'_{sl})$ are defined as below

$$\mathbf{S}_{sl} = \frac{\sum_{t=1}^{T}(o_t - \mu_{sl})(o_t - \mu_{sl})'\gamma_{slt}}{\gamma_{sl}}$$

$$Q(\mathbf{C}_{sl}, \mathbf{C}'_{sl}) = \log|\mathbf{C}_{sl}| - tr(\mathbf{S}_{sl}\mathbf{C}_{sl}) - \frac{\lambda_{sl}}{\gamma_{sl}}||\mathbf{C}_{sl}||_1.$$

From equation (5), we can see that maximizing $Q(\mathbf{C}, \mathbf{C}')$ is equivalent to maximizing each individual $Q(\mathbf{C}_{sl}, \mathbf{C}'_{sl})$. Because λ_{sl} is a user defined hyperparameter, finding $\hat{\mathbf{C}}_{sl}$ that maximize $Q(\mathbf{C}_{sl}, \mathbf{C}'_{sl})$ is equivalent to finding $\hat{\mathbf{C}}_{sl}$ that satisfies

$$\hat{\mathbf{C}}_{sl} = \underset{\mathbf{C}_{sl}}{\operatorname{argmax}}\{\log|\mathbf{C}_{sl}| - tr(\mathbf{S}_{sl}\mathbf{C}_{sl}) - \lambda_{sl}||\mathbf{C}_{sl}||_1\}. \quad (6)$$

This can be solved by using the *graphic lasso* program[6]. In this paper we only investigate global penalization (i.e. one λ for all Gaussian components in equation (6)).

It is worth mentioning that we can even modify the objective function to penalize each element in the inverse covariance matrices differently. Then the optimal $\mathbf{C}_{sl}$s can be found by using

$$\hat{\mathbf{C}_{sl}} = \underset{\mathbf{C}_{sl}}{\operatorname{argmax}}\{\log|\mathbf{C}_{sl}| - tr(\mathbf{S}_{sl}\mathbf{C}_{sl}) - ||\mathbf{C}_{sl} * \mathbf{P}_{sl}||_1\},$$

where $*$ indicates componentwise multiplication and $\mathbf{P}_{sl} = \{\lambda_{ij}\}_{sl}$ with $\lambda_{ij} = \lambda_{ji}$ is a user defined penalization matrix. This can also be solved by using the *graphic lasso* program. When all the elements inside $\mathbf{P}_{sl}$ are zero, this is equivalent to training a full covariance system. When the diagonal elements of $\mathbf{P}_{sl}$ are set to zero while others are set to infinity, the system trained is a diagonal covariance system.

3. EXPERIMENTAL SETUP AND RESULTS

The SI-84 database (about 14.5 hours) from WSJ0 was used as our training material. To see how the performance was affected by the size of the training data, we also randomly sampled 1 hour, 3 hours, 5 hours and 10 hours of data from the SI-84 database. The Nov'92 data was used as the evaluation data set. The audio data was represented by $d = 39$ feature vectors every 10ms: 12MFCC plus c_0 with ceptral mean substraction and delta and acceleration coefficients.

We have run experiments using the standard bigram language model provided by the LDC. The pronunciation information came from the CMU dictionary. Cross-word triphone models were used and all HMMs had three emitting states and a strictly left-to-right topology. The HTK tool was used in our experiments [9].

3.1. Performance affected by varying λs

To see how the system performance was affected by different λs, we first trained systems with one mixture component per state and with different λs using the randomly sampled one hour database. The word error rates (WER) on the evaluation set are shown in Figure 1. The results are plotted against the percentage of nonzero parameters relative to the full covariance matrices (average number of nonzero parameters in the sparse upper triangular matrix divided by d*(d+1)/2). The corresponding λs are also given. As can be seen from the plot, with λ increasing, more parameters of the inverse covariance matrices are driven to zero. The left-most point on the plot corresponds to a system with diagonal covariance matrices. The right-most point corresponds to a system with full covariance matrices.

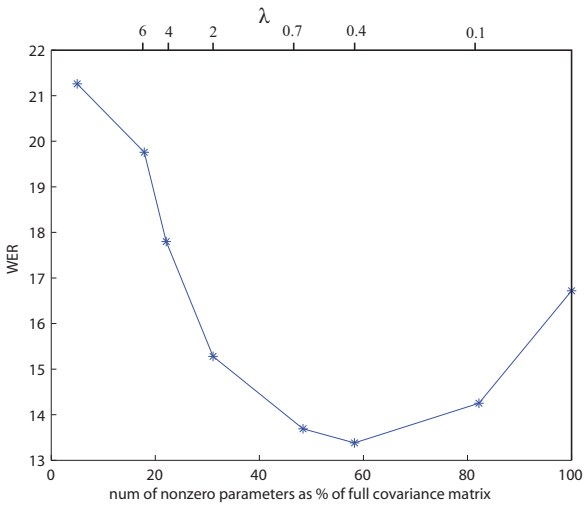

Fig. 1. Word error rate on the evaluation data with varying λs.

From this plot, we can see that sparse inverse covariance systems can substantially outperform full covariance systems. With λ equal to 0.4, the absolute improvement is 3.6%. It also shows that only about 25% of the parameters of the inverse covariance matrices are needed to achieve the same performance of a full covariance system.

3.2. Performance affected by training data size

We also investigated the performance affected by the size of the training data. We trained a diagonal covariance system on the SI-84 database that got its peak performance with 8 mixture components per state (624 Gaussian parameters per state). To avoid vast increase of free parameters, all the systems described in this section are trained with 2 mixture components per state (1638 Gaussian parameters per state for full covariance systems). In the above experiments, we found that setting λ to 0.4 leads to good performance. Thus in the following experiments, λ was always set to 0.4. We trained systems with different amount of training data. The results are showed in Figure 2. When the training data is very sparse (e.g. 3 hour), the diagonal covariance systems seems better than the full covariance systems. However, as more training data is available, the full covariance systems outperform the diagonal covariance systems with the same number of mixture components per state. The plot also shows that the sparse inverse covariance systems consistently work better than the full covariance and the diagonal covariance systems. We also

found that when the training data is sparse (e.g. 3 hours and 5 hours) the absolute improvement achieved by using sparse inverse covariance matrices is much more, compared with the performance improvement achieved when more training data is available. This is because when the training data is sparse the full covariance systems tend to be over-fitting and generalize poorly to unseen test data. By adding a penalization term to the objective function to penalize complex models, the systems generalize much better. Using all the SI-84 data, the word error rate of the system with sparse covariance matrices with only two mixture components per state is 5.94%. This is even better than the best performance (WER:6.17%) that we can get with a diagonal covariance system with 8 mixture component per state trained on the same data.

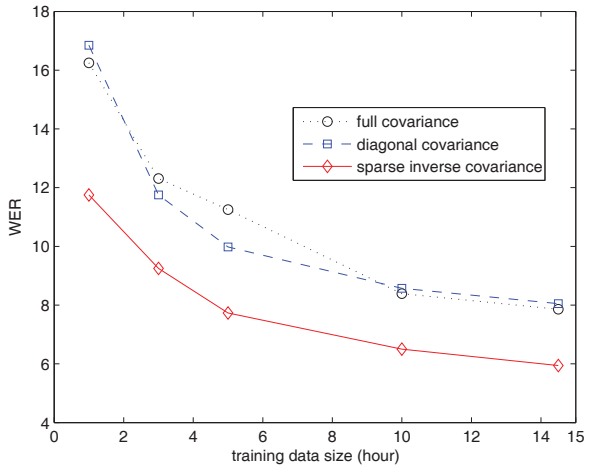

Fig. 2. Comparisons of systems with diagonal, full and sparse inverse covariance matrices with different amount of training data. All λs are set to 0.4 during the training of sparse inverse covariance systems.

4. CONCLUSION AND FUTURE WORK

In this paper, We address the problem of low resource acoustic model training by considering sparse inverse covariance matrices. We propose to add an l_1 penalization term to the traditional objective function for MLE in order to automatically drive the inverse covariance matrices to sparse. The *graphic lasso* algorithm is incorporated into the EM procedure to learn parameters of HMM using the new objective function.

The experimental results show that the sparse inverse covariance systems consistently work much better than the full covariance systems, especially when the training data is sparse. Our results also show that only about 25% of the parameters of the inverse covariance matrices are needed to achieve the same performance of full covariance systems. Since many of the elements inside the inverse covariance matrices are driven to zero, much less computation is needed.

In the future, we would like to investigate how to improve the computation and memory efficiency of acoustic models by using sparse inverse covariance matrices. We will also investigate different penalization methods, for example, the value of λs depend on the size of data available for the particular state or to smooth off the diagonal as proposed in [10].

5. ACKNOWLEDGEMENT

We would like to thank the Hong Kong Research Grants Council (RGC) for partially supporting this work through Hong Kong PhD fellowship scheme.

6. REFERENCES

[1] D. Povey et. al., "Subspace gaussian mixture models for speech recognition," in *ICASSP*. IEEE, 2010, pp. 4330–4333.

[2] S. Axelrod, V. Goel, R.A. Gopinath, P.A. Olsen, and K. Visweswariah, "Subspace constrained gaussian mixture models for speech recognition," *IEEE Transactions on, Speech and Audio Processing*, vol. 13, no. 6, pp. 1144–1160, 2005.

[3] SS Chen and R.A. Gopinath, "Model selection in acoustic modeling," in *Proc. of Eurospeech*. Citeseer, 1999, vol. 3, pp. 1087–1090.

[4] J.A. Bilmes, "Factored sparse inverse covariance matrices," in *ICASSP*. IEEE, 2000, vol. 2, pp. II1009–II1012.

[5] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *The Journal of Machine Learning Research*, vol. 9, pp. 485–516, 2008.

[6] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432, 2008.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[8] J.A. Bilmes, "A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *International Computer Science Institute*, vol. 4, pp. 126, 1998.

[9] S.J. Young et. al., "The htk book version 3.4," 2006.

[10] D. Povey, "Spam and full covariance for speech recognition," in *Ninth International Conference on Spoken Language Processing*, 2006.