Sparse Gaussian Graphical Models for Speech Recognition

Peter Bell, Simon King

Centre for Speech Technology Research, University of Edinburgh, UK

peter.bell@ed.ac.uk, simon.king@ed.ac.uk

Abstract

We address the problem of learning the structure of Gaussian graphical models for use in automatic speech recognition, a means of controlling the form of the inverse covariance matrices of such systems. With particular focus on data sparsity issues, we implement a method for imposing graphical model structure on a Gaussian mixture system, using a convex optimisation technique to maximise a penalised likelihood expression. The results of initial experiments on a phone recognition task show a performance improvement over an equivalent full-covariance system.

Index Terms: speech recognition, acoustic models, graphical models, precision matrix models

1. Introduction

Most modern systems for automatic speech recognition (ASR) use hidden Markov models (HMMs) with the the acoustic feature observation probabilities modelled by a mixture of multivariate Gaussian components, known as a Gaussian mixture model (GMM). An important consideration is the form of each Gaussian component. Commonly, individual acoustic features are assumed uncorrelated within each component, and so the Gaussian is taken to have a diagonal covariance matrix. This assumption can be made in practice because of the use of front-end feature decorrelation techniques, and also because a full-covariance Gaussian component can be modelled by splitting into multiple diagonal-covariance components.

However, these techniques are not fully effective at removing the need to model correlations, and recent research has focused on finding alternative forms for the covariance matrix. It is desirable to find a compromise between the need for just p covariance parameters, in the diagonal case, and $\frac{1}{2}p(p+1)$ parameters for a full covariance matrix (where p is the number of acoustic features for each frame). In the latter case, much larger amounts of training data are required to obtain good parameter estimates: otherwise the model is likely to be over-fitted to the training data, leading to poor recognition performance. The problem becomes particularly acute if it is desired to increase the number of acoustic features used (for example, by adding information from surrounding frames to the feature vector).

[1] proposed fitting the covariance structure to correspond to a pre-determined sparse graphical model. Such models encode information about the *conditional* independence structure of the system, and are described further in Section 2. When applied to a multivariate Gaussian system, the model is known as a Gaussian graphical model (also a covariance selection model, after [2]). It has been argued that graphical models can provide a more parsimonious statistical model for speech [3]. It will be shown that adopting a given graphical model for the system imposes constraints on the sparsity structure of the inverse covariance matrix, known as the *precision matrix*. A benefit of modelling the structure of this matrix, rather than the covariance matrix, is that it is used directly in the Gaussian probability calculation, yielding computational benefits for speech decoding. Indeed, much other work has been undertaken to reduce the dimensionality of the precision matrix. A review is given in [4].

This paper considers the problem of structure learning in Gaussian graphical models, and coupled with it the problem of parameter estimation, given the structural constraints. As we explain in Section 2.2, other approaches to this problem used in ASR systems often suffer from sparse data issues. We investigate the use of a technique recently introduced by [5], related to methods based on the Lasso [6]. These methods seek to estimate the graphical model structure by maximising a penalised likelihood expression, and have been tested on inference problems in Bioinformatics. The penalty term is chosen so that the maximisation results in a sparse precision matrix, whilst ensuring that an optimal solution to the maximisation problem can be efficiently found. We describe the procedure in detail in Section 3. We adapt this technique for learning a set of GMM-based acoustic models with reduced dimensionality, and test it on a standard phone recognition task. In doing this we consider the issue of how to chose the appropriate size of penalty term.

2. Background

2.1. Gaussian Graphical Models

In its undirected form, a graphical model [7] for random variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ (corresponding to acoustic features) [1] consists of a graph, $\mathcal{G} = (V, E)$, with vertices $V = \{1, \ldots, p\}$ representing random variables and undirected arcs $E \subseteq V \times V$ representing causal relationships between variables. The absence of an edge (i, j) indicates that X_i and X_j are (pairwise) conditionally independent, given the other variables in the system:

$$X_i \perp\!\!\!\perp X_j \,|\, X \backslash \{X_i, X_j\} \tag{1}$$

which is equivalent to the distribution function admitting the factorisation

$$f(\mathbf{x}) = f(x_i, x_s) f(x_j, x_s) \tag{2}$$

where $x_s = \{x_l : l \neq i, l \neq j\}$. The distribution function of a mean-centred[2] multivariate Gaussian distribution with precision matrix $\Omega = \Sigma^{-1}$ is given by

$$f(\mathbf{x}; \Omega) = (2\pi)^{-\frac{p}{2}} (\det \Omega)^{\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{x}^T \Omega \mathbf{x}^T)$$

[1] We use bold type to denote vectors of random variables or observations.

[2] We assume throughout that the means of each distribution can be reliably estimated, and take all data to be mean-centred for the sake of clarity.

and from this it is clear that X_i and X_j are conditionally independent if and only if $\Omega_{ij} = 0$.

Dempster [2] derived the maximum likelihood estimator for the covariance matrix of the restricted distribution, $\Sigma_{\mathcal{G}}$, showing that we must have it matching the sample covariance matrix, S, whenever there is a corresponding arc in $\mathcal{G}$:

$$[\Sigma_{\mathcal{G}}]_{ij} = s_{ij} \qquad (i,j) \in E \tag{3}$$

and it can be shown that finding such a matrix with its inverse matching the graphical model is equivalent to finding the unique positive definite matrix $\Sigma_{\mathcal{G}}$ with maximum determinant, subject to the constraints (3).

2.2. Structure learning

Determining the structure of $\mathcal{G}$ – that is, fixing the edge set E or the sparsity structure of Ω – is an important problem. It is possible to specify this according to prior knowledge: fixing Ω to be block diagonal or banded diagonal would be an example. However, we would expect that the optimal structure could be learnt from the available data, and research has been undertaken in this area for graphical models in general [8, 9]. In the case of Gaussian graphical models, the structure can be tied across different Gaussians (for example, between mixture components of the same state, or across triphones) by factorising the precision matrix [10] according to

$$\Omega = LDL^T \tag{4}$$

and tying the lower triangular matrix L across states. D is a state-specific diagonal matrix of feature variance. [1] selected edges based on estimates of class-conditional Mutual Information (MI) between features, with the aim of creating discriminatively structured models. Models selected according to these criteria were shown to outperform graphical models with the same number of randomly-selected arcs. However, an critical deficiency of MI-based selection is that MI is essentially a function of feature correlation, and so a reliable estimate of the correlation is required.

This is problematic: correlation values are obtained from the sample covariance matrix, which is often a poor estimate of the true covariance, particularly in the case that the number of available samples, n, is close to the number of features, p, – and it is not even guaranteed to be positive definite in the case $n < p$. Even when the true graphical model structure is highly sparse, with few parameters required to be estimated, it is not simple to recover this structure from the inverted sample covariance matrix due to the statistical "noise" present in the estimate of the (generally non-sparse) full covariance matrix.

One solution is to use a backward selection procedure: the importance of each edge (according to the chosen measure) is computed using the parameters learnt for a full-covariance system. The least important edges are removed one at a time from the graphical model. After each step, the covariance parameters are then recomputed using the new model. However, this greedy search does not guarantee a globally optimal solution, and has the disadvantage that the parameter re-estimation after each step imposes a high computational cost.

3. Convex Optimisation Techniques

3.1. Penalised Likelihood

A standard technique in statistical modelling to avoid overfitting to data is to find the parameter set maximising a penalised likelihood expression, where the size of penalty term increases with the number of free parameters in the model. An well-known example is the Bayesian Information Criterion [11].

Recent work [6, 5, 12] on Gaussian graphical model structure learning has focused on the use of the l_1-norm of the parameter vector as a penalty term. The l_q-norm of a parameter vector $\theta = (\theta_1, \ldots, \theta_p)$ is given by $||\theta||_q^p = \sum_i |\theta_i|^q$. As noted in [6], the advantage of taking $q = 1$ is that it is the unique choice both giving a sparse parameter set (which occurs only for $q \leq 1$), and resulting in the maximisation problem θ being convex [13] (which occurs only for $q \geq 1$) and hence efficiently soluble.

We follow the approach of [5]. Here, the sparse precision matrix, Ω, is found by maximising the penalised log likelihood of the data. The penalty term is the l_1 norm of all conditional correlation parameters, given by the sum of the magnitude of the off-diagonal elements of Ω. Ignoring constant terms, the log likelihood of Ω, given data $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is

$$\ell_N(\Omega) = \frac{N}{2}\log\det\Omega - \sum_n^N \frac{1}{2}\mathbf{x}_n^T\Omega\mathbf{x}_n \tag{5}$$

$$= \frac{N}{2}\log\det\Omega - \mathrm{tr}\sum_n^N \frac{1}{2}\Omega\mathbf{x}_n^T\mathbf{x}_n \tag{6}$$

$$= \frac{N}{2}(\log\det\Omega - \mathrm{tr}\,\Omega S) \tag{7}$$

where we use the fact that the trace operator is invariant to cyclic permutations, and S denotes the covariance matrix of the data. We therefore obtain the sparse version of Ω by solving

$$\hat{\Omega} = \arg\max_{\Omega \succ 0}\{\ell_N(\Omega) - \frac{\rho N}{2}||\Omega||_1\} \tag{8}$$

$$= \arg\max_{\Omega \succ 0}\{\log\det\Omega - \mathrm{tr}\,\Omega S - \rho||\Omega||_1\} \tag{9}$$

The parameter ρ is used to control the size of the penalty term (with an increased penalty term leading to a more sparse matrix). Both the parameter set and graphical model structure are obtained simultaneously.

Details of an iterative algorithm for solving this problem can be found in [5]. $\hat{\Omega}$ is guaranteed to be positive definite after each iteration.

3.2. Penalised likelihood for Gaussian Mixture Models

The method described above can be used for structure learning in a GMM with no difficulty. The density function of a GMM is given by

$$g(\mathbf{x}_n; \theta) = \sum_k w_k f(\mathbf{x}_n; \Omega_k) \tag{10}$$

where w_k is the weight of the Gaussian component k (weights summing to 1) and $f(\mathbf{x}; \Omega_k)$ is the density function of the corresponding Gaussian. θ denotes the complete parameter set.

No analytical solution exists for determining the maximum likelihood estimates of these parameters, so an iterative method is required. Suppose that to each $\mathbf{x}_n$ we attach weights $p(k|n)$, denoting the probability that x_n has been generated by component k. For the log likelihood we have

$$\ell(\mathbf{x}_n; \theta) = \log g(\mathbf{x}_n; \theta) \tag{11}$$

$$= \log \sum_k p(k|n)\frac{w_k f(\mathbf{x}_n; \Omega_k^{(i)})}{p(k|n)} \tag{12}$$

Given a set of initial parameter estimates, denoted $w_k^{(i)}$, $\Omega_k^{(i)}$, it can be shown that a lower bound for the likelihood of the model is given by

$$\sum_k p^{(i)}(k|n)\log(w_k f(\mathbf{x};\Omega_k^{(i)})) \tag{13}$$

where

$$p^{(i)}(k|n) = \frac{w_k f(\mathbf{x};\Omega_k^{(i)})}{\sum_m w_m f(\mathbf{x};\Omega_m^{(i)})} \tag{14}$$

and consequently at each iteration we choose parameters $w_k^{(i+1)}$, $\Omega_k^{(i+1)}$ to maximise this expression, summed for all data points. This is the familiar Expectation Maximisation (EM) algorithm [14].

Suppose we now impose a penalty term on the log likelihood for $\mathbf{x}_n$:

$$\sum_k \frac{\rho}{2} p(k|n)||\Omega_k||_1 \tag{15}$$

weighting each penalty according to the probability of the point belonging to component k. Since this is not a function of the data, it does not affect the form of the lower bound (including the computation of $p^{(i)}(k|n)$), and so we maximise at each iteration, the value of

$$\sum_n \sum_k p^{(i)}(k|n)(\log(w_k f(\mathbf{x};\Omega_k^{(i)})) - \frac{\rho}{2}||\Omega_k||_1) \tag{16}$$

Considering just terms in Ω_k, we maximise

$$\sum_n p(k|n)^{(i)}(\log\det\Omega_k - \rho||\Omega_k||_1 - \mathrm{tr}\,\Omega_k\mathbf{x}^T\mathbf{x}) \tag{17}$$

$$\begin{aligned} = &\sum_n p(k|n))^{(i)}(\log\det\Omega_k - \rho||\Omega_k||_1) \\ &- \mathrm{tr}\,\Omega_k \sum_n p(k|n)^{(i)}\mathbf{x}^T\mathbf{x} \end{aligned} \tag{18}$$

from which it can be seen that we simply carry out the same optimisation (9) as before, setting the sample covariance matrix

$$S^{(i)} = \frac{\sum_n p(k|n)^{(i)}\mathbf{x}_n^T\mathbf{x}_n}{\sum_n p(k|n)^{(i)}} \tag{19}$$

at each iteration.

3.3. Choosing the penalty parameter

We consider the choice of the parameter ρ. As the sample size tends to infinity, it is known that the unpenalised maximum-likelihood estimate is optimal, so we should have $\rho(n) \to 0$. It was observed empirically in [5], using synthetic data, that the best fit to the true graphical model structure was obtained by setting ρ equal to the standard deviation of the sample covariance matrix, approximately proportional to $n^{-0.5}$. Our own preliminary experiments on synthetic data sets of varying sizes appeared to confirm that this choice is optimal for classification: however, we have yet to carry out exhaustive experiments to determine whether or not this is true for speech decoding, and in this work use a cross-validation method.

4. Experiments

To investigate the potential benefits for ASR of the penalised likelihood structure learning technique, we carried out phone recognition experiments on the TIMIT corpus. A standard monophone HMM baseline system was constructed using 48 phone models, each with three emitting states. The acoustic feature vector consisted of 12 MFCCs plus energy component, their deltas and double-deltas. The observation probability for each state was modelled by a diagonal covariance GMM, with successive model sets having increasing numbers of Gaussian components (the same for each phone). Results were obtained on the reduced test set of 192 utterances, collapsing the labels to the usual 39-phone set. A phone-level bigram language model was used for decoding: early experiments indicated that applying a scaling factor to the language model of 5.0, and using an insertion penalty of 2.5, tended to give highest accuracy scores, and these were fixed for all subsequent experiments. The baseline scores were very close to figures from similar systems presented in other work. The trend is for recognition performance to improve up to about 48 components per state.

In obtaining full-covariance statistics for GMMs using the EM algorithm, it is necessary to consider the choice of initial parameters. We found that systems initialised from a diagonal covariance GMM with the same number of Gaussian components performed much better than those initialised by component-splitting a full-covariance system with a smaller number of components. (An alternative, initialising from a sparse graphical model system with a smaller number of parameters, was judged to be too complex at this stage, given the need to experiment with the size of the penalty parameter at each split). To reduce computational overhead when re-estimating the parameters, re-estimation was carried out independently for each model using a forced alignment of the training data, generated using the 24-component diagonal covariance system. No parameters were shared between models or components.

Whilst it may be true that sparse graphical model GMMs are effective with very high numbers of components, here experiments were carried out on full-covariance systems of 12 and 16 components. These gave performance comparable with the best diagonal-covariance systems, whilst tending to avoid problems caused by non-invertible sample covariance matrices (this occurs more frequently as the number of components is increased). Before performing the optimisation (9) to obtain the sparse precision matrices from the accumulated sample covariance statistics, the sample matrices were scaled so that all partial sample correlations were equal to one.

Figure 1 shows the variation in recognition performance with the penalty parameter ρ for models with 12 sparse-precision components, with the equivalent full-covariance performance corresponding to $\rho = 0$. Table 1 shows a phone accuracy results for a range of model sets: for the diagonal covariance baseline systems and full-covariance systems with varying numbers of Gaussian components, and for sparse precision matrix systems with varying penalty size. It can be seen that the performance of the optimum sparse precision matrix system achieves the highest accuracy score of any system. It outperforms a full-covariance system with the same number of mixture components by 0.7% absolute. Although the increase is small, it was found to be statistically significant at the 1% level by comparing scores on a per-speaker basis – and 163,000 fewer parameters are used.

5. Discussion

We have demonstrated that learning Gaussian mixture models with sparse precision matrices may be beneficial for ASR tasks. However, further investigation is needed into the the models' performance on more challenging tasks.

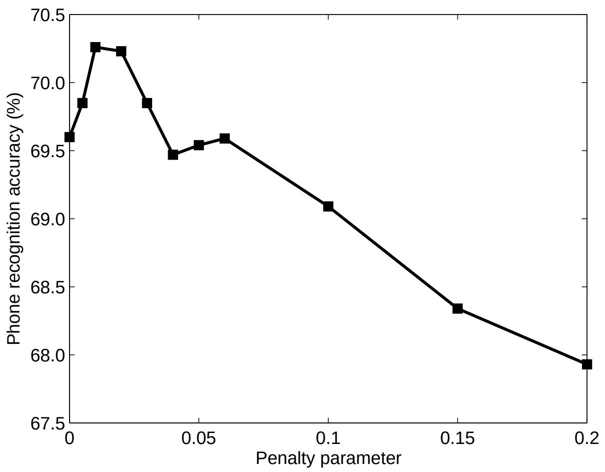

Figure 1: *Phone accuracy of the sparse graphical model systems with varying penalty parameter.*

Table 1: *Phone accuracy scores. Parameter counts include only precision matrix parameters.*

Type	#GCs	#Params	Accuracy
Diag	1	5,616	57.6
Diag	12	67,392	67.0
Diag	24	134,784	68.7
Diag	36	202,176	69.0
Diag	48	269,568	68.9
Full	1	112,320	63.6
Full	2	224,640	65.6
Full	12	1,347,840	**69.6**
Sparse	12	1,234,536	69.9
Sparse	12	1,184,787	**70.3**
Sparse	12	1,108,465	70.2
Sparse	12	849,211	69.6
Sparse	12	697,590	69.1
Sparse	12	453,261	67.9

There are a number of issues that we hope to address in future work. Firstly, we have observed that the precision matrix structure learning is effective only if the matrices are first scaled so that all partial correlations are equal to one. This is not possible if the covariance matrix is singular, which is increasingly likely to occur as the number of Gaussian components is increased. This lead to us considering only models with 12 components or fewer, for which covariance matrices can be reliably estimated, limiting the potential benefit to be gained by the penalised likelihood method – particularly the benefit that it always returns a non-singular solution. It will be necessary to modify the algorithm to implement this scaling directly as part of the optimisation procedure. We would hope to see greater improvements over full-covariance models in sparse data conditions or with large numbers of mixture components: neither could be fully investigated here.

Secondly, we will need to conduct further experiments to determine any possible analytic form of the optimal penalty parameter, which would include controlling the amount of data available for training each model, rather than simply using a standard training set for all experiments. Tying precision matrices across components will allow these large-scale experiments to be carried out more efficiently and is a standard method in precision matrix modelling.

Thirdly, no attempt had been made here to learn graphical model structures in a discriminative fashion, focusing instead on the benefits of penalised likelihood from a purely statistical viewpoint. We will aim to modify the technique to learn a discriminatively beneficial structure. Other precision matrix modelling methods [4] have trained models according to the minimum phone error criterion (MPE), whilst [1] used an explicitly discriminative measure of the significance of each edge in the graphical model. The challenge will be to incorporate aspects of these techniques, whilst retaining the attractive properties of the current estimator.

6. References

[1] J. Bilmes, "Factored sparse inverse covariance matrices," in *Proceedings of ICASSP*, Istanbul, Turkey, June 2000.

[2] A. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[3] J. A. Bilmes, *Mathematical Foundations of Speech and Language Processing*. Institute of Mathematical Analysis, 2003, ch. Graphical Models and Automatic Speech Recognition.

[4] K. Sim and M.J.F.Gales, "Precision matrix modelling for large vocabulary continuous speech recognition," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR.485, June 2004.

[5] O. Banerjee, A. d'Aspremont, and L. E. Ghaoui, "Convex optimization techniques for fitting sparse gaussian graphical models," in *Proceedings of ICML*, Pittsburgh, PA, June 2006.

[6] N. Meinshausen and P. Bühlman, "High dimensional graphs and variable selection with the Lasso," *Annals of Statistics*, vol. 34, pp. 1436–1462, 2006.

[7] S. L. Lauritzen, *Graphical Models*. Oxford University Press, 1996.

[8] M. Deviren and K. Daoudi, "Continuous speech recognition using structural learning of dynamic Bayesian networks," in *Proceedings of EUSIPCO*, Toulouse, France, Sept. 2002.

[9] N. Friedman, K. Murphy, and S. Russell, "Learning the structure of dynamic probabilistic networks," in *Proceedings of UAI*, Madison, WI, July 1998.

[10] M. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, May 1999.

[11] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[12] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 2005, To appear.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[14] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.