interspeech 2008 incorporating SST 08

A Shrinkage Estimator for Speech Recognition with Full Covariance HMMs

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 consider the problem of parameter estimation in full-covariance Gaussian mixture systems for automatic speech recognition. Due to the high dimensionality of the acoustic feature vector, the standard sample covariance matrix has a high variance and is often poorly-conditioned when the amount of training data is limited. We explain how the use of a shrinkage estimator can solve these problems, and derive a formula for the optimal shrinkage intensity. We present results of experiments on a phone recognition task, showing that the estimator gives a performance improvement over a standard full-covariance system.

Index Terms: speech recognition, acoustic models, regularisation, full covariance estimation, shrinkage estimator

10.21437/Interspeech.2008-106

1. Introduction

HMM-based systems for automatic speech recognition (ASR) typically model the acoustic features using mixtures of multivariate Gaussians (GMMs). A variety of schemes have been proposed for controlling the number of covariance parameters of each Gaussian, which may vary between p, in the diagonal case, and $\frac{1}{2}p(p+1)$ in the full covariance case, where p is the size of the acoustic feature vector. Examples include Semi-Tied Covariance Matrices (STC) [1], Extended Maximum Likelihood Linear Transforms [2] and Subspace for Precision and Mean [3]. Studies have shown that in general, the greater the number of parameters used in the models, the better the recognition performance [4].

These results have lead us to consider the case where each Gaussian is modelled with the full $\frac{1}{2}p(p+1)$ parameters. Given observations $\mathbf{x}(t)$, with probabilities $\gamma_m(t)$ of being generated by a Gaussian m, the maximum likelihood estimate of the true covariance matrix, Σ_m is given by

$$S_m = \frac{\sum_t \gamma_m(t)(\mathbf{x}(t) - \hat{\mu}_m)(\mathbf{x}(t) - \hat{\mu}_m)^T}{\sum_t \gamma_m(t)} \qquad (1)$$

where $\hat{\mu}_m$ is the maximum likelihood estimate of the mean for the Gaussian m. We set $\beta_m = \sum_t \gamma_m(t)$.

The maximum likelihood estimator (MLE) has several attractive properties. Firstly, it is *consistent*: the MLE of a parameter θ, based on n samples, tends to θ as $n \to \infty$. Secondly, it is *asymptotically efficient*: as $n \to \infty$, the MLE has the minimum variance achieved by any unbiased estimator. Furthermore, the maximum likelihood estimate is easy to compute from data. In HMM-GMM systems, whilst the MLE cannot be obtained directly, the well-known EM algorithm is an iterative algorithm, with each successive estimate guaranteed to increase the likelihood of the training data.

However, use of ML estimation for machine learning has several drawbacks. Learning theorists have challenged the usefulness of asymptotic results, pointing out that rarely can we consider the amount of training data to be approaching infinity [5]. In ASR applications where high-dimensional feature vectors are usual, the amount of training data may not even be much larger than the number of parameters required to be estimated. The problem is particularly acute for full covariance matrix estimation, with $\mathcal{O}(p^2)$ parameters. Furthermore, in limited data situations, the ML covariance matrix estimate is likely to be ill-conditioned: that is to say, the ratio between the largest and smallest eigenvalues is large, resulting in the amplification of numerical errors when the matrix is inverted – as it is when computing the probability density function of the Gaussian. In the extreme case, when $n < p$, the matrix is, in general, non-invertible.

In this paper we aim to address two problems associated with limited data covariance-matrix estimation in ASR. Firstly, the problem that the MLE does *not* have minimum variance in the case that the asymptotic assumption does not hold. If an estimator has high variance then it is likely to be over-fitted to the training data, leading to poor performance on test data. Secondly, the problem of the matrix being ill-conditioned. Both problems can be solved by regularising the estimator appropriately.

This paper is structured as follows: firstly we introduce the shrinkage estimator – as a regularised covariance matrix estimator – and discuss its properties. We briefly discuss other methods for covariance regularisation. We go on to explain how the shrinkage estimator can be obtained in the context of an HMM-GMM system, and test the new estimator on a phone recognition task where the amount of training data is limited.

2. The shrinkage estimator

2.1. Introducing shrinkage

Stein [6] first introduced the concept of "shrinkage" as applied to high-dimensional estimators (specifically, of the mean of a distribution), deriving the surprising result that the performance of the MLE can always be improved upon by shrinking by a given factor λ (the "shrinkage intensity"). More recently, Ledoit and Wolf [7] showed how this procedure can be applied to covariance matrices. We consider a new estimator, U, of Σ, given by

$$U = (1 - \lambda)S + \lambda D \qquad (2)$$

where D, the "shrinkage target", is a diagonal matrix. It can be seen that as λ is increased to one, the off-diagonal elements of U shrink towards zero. (For clarity, we have suppressed the dependence on m).

Adopting the terminology of classical statistics, we measure the performance of an estimator X of a parameter θ by its mean squared error (MSE), given by

$$\text{MSE}(X) = \mathbb{E}(X - \theta)^2 \qquad (3)$$

Accepted after peer review of full paper

(where the expectation is with respect to θ). Note the standard result, that

$$\text{MSE}(X) = \mathbb{E}((X - \mathbb{E}X) + (\mathbb{E}X - \theta))^2 \quad (4)$$
$$= \mathbb{E}(X - \mathbb{E}X)^2 + (\mathbb{E}X - \theta)^2 \quad (5)$$
$$= \text{var}(X) + \text{bias}^2(X) \quad (6)$$

Typically, a higher dimensional estimator will have a lower bias, but higher variance – minimising the MSE of an estimator can be viewed as optimising the trade-off between the two.

Subject to a minor correction factor, S is an unbiased estimator of Σ, whilst D is biased in its off-diagonal elements. The shrinkage procedure can therefore be viewed as "backing off" from the high-variance, unbiased S to the low-variance, biased D. It will be seen below that the optimal shrinkage factor λ can be directly computed.

To measure the MSE of a matrix estimator, we use the Frobenious norm, given by

$$\|A\|_F = \sqrt{trA^TA} = (\sum_i \sum_j |A_{ij}|^2)^{\frac{1}{2}} \quad (7)$$

This arises from the inner product $\langle A, B\rangle = \text{tr}\, A^T B$. The MSE of U is given by $\mathbb{E}\|U - \Sigma\|_F^2$. In the equations that follow, the Frobenius norm is used implicitly.

In [7], D is taken to be a uniform diagonal matrix $D = \rho I$. However, Schäfer and Strimmer [8] discuss a variety of alternative targets. As they explain, the case where D consists of the diagonal elements of S is attractive for several reasons, and it is this target which we use throughout this work.

2.2. Finding the optimal shrinkage intensity

In [7] a method was obtained for computing the optimal shrinkage intensity analytically, whilst [8] generalised them to a variety of shrinkage targets. We seek λ to minimise

$$\mathbb{E}\|U - \Sigma\|^2 = \mathbb{E}\|\lambda(D - \Sigma) + (1 - \lambda)(S - \Sigma)\|^2 \quad (8)$$
$$= \lambda^2\mathbb{E}\|D - \Sigma\|^2 + (1 - \lambda)^2\mathbb{E}\|S - \Sigma\|^2 + 2\lambda(1 - \lambda)\mathbb{E}\langle D - \Sigma, S - \Sigma\rangle \quad (9)$$

Differentiating with respect to λ and setting the result equal to zero, we obtain

$$\lambda[\mathbb{E}\|D - \Sigma\|^2 + \mathbb{E}\|S - \Sigma\|^2 - 2\mathbb{E}\langle D - \Sigma, S - \Sigma\rangle] = \mathbb{E}\|S - \Sigma\|^2 - \mathbb{E}\langle D - \Sigma, S - \Sigma\rangle \quad (10)$$

$$\lambda\mathbb{E}\|(S-\Sigma)-(D-\Sigma)\|^2 = \mathbb{E}\langle S-\Sigma, (S-\Sigma)-(D-\Sigma)\rangle \quad (11)$$

and so

$$\lambda = \frac{\mathbb{E}\langle S - \Sigma, S - D\rangle}{\mathbb{E}\|S - D\|^2} \quad (12)$$

In the case that D consists of the diagonal elements of S, and using the fact that S is unbiased, the numerator becomes the sum of off-diagonal elements of the matrix $\text{var}(S)$. The denominator becomes the expected sum of the off-diagonal elements of S. From the formula (12) it can be noted that λ increases with $\text{var}(S)$, so that for small sample sizes, the shrinkage target, D, achieves more prominence.

2.3. Estimator properties

2.3.1. Matrix conditioning

In [7] it is shown that the eigenvalues of the sample covariance matrix are, on average, more dispersed than the eigenvalues of the true covariance matrix. This means that the sample covariance matrix is likely to be less well-conditioned that the true covariance matrix, leading to numerical problems when the matrix is inverted. The eigenvalues of D (just the diagonal elements) have the same mean as the eigenvalues of S, but are less dispersed, so the linear combination of S and D used in the shrinkage estimator will shrink the eigenvalues towards m. (See [7] for technical probabilistic results concerning the conditioning of the shrinkage estimator).

2.3.2. Bayesian viewpoint

For simplicity of presentation, we have not adopted a Bayesian framework in this paper. However, it is quite possible to consider the above equations in this way. The MSE of an estimator U is equivalent to the Bayes' risk with a quadratic loss function:

$$R(U) = \int_\Sigma \|U - \Sigma\|^2 f(\Sigma|\mathbf{x})d\Sigma \quad (13)$$

where $f(\Sigma|\mathbf{x})$ is the posterior probability. If a non-informative prior is chosen, we obtain the minimum risk at $U = S$ as in the classical MLE case. The shrinkage estimator can be obtained by choosing a suitable prior for Σ centred on the shrinkage target. As the amount of data is reduced, the influence of the prior is increased, and the minimum Bayes' risk estimator becomes closer to D.

2.4. Alternative estimators

Most large-vocabulary ASR systems employ some form of covariance regularisation. Most simply, flooring diagonal covariance elements to some proportion of global variance is standard practice, considered essential in systems with many Gaussians. When p is large, the covariance matrix may be constrained to have a block-diagonal structure: in this case, the minimum number of samples required for the sample matrix to be invertible is equal to the size of the largest block.

Methods such as [1, 2, 3], do not set out to regularise the covariance matrices, but the sharing of covariance parameters across Gaussians does reduce the variance of the estimators used. The full covariance systems described in [9] apply smoothing to off-diagonal elements of the covariance matrices. In effect this results in a shrinkage estimator: however, the smoothing factor used does not have the same optimality properties as described here, and requires a hyper-parameter to be specified. In the same work, smoothing functions are used to control the minimum eigenvalues of discriminatively-estimated covariance matrices. The maximum eigenvalues are not controlled, however, so the matrices are not necessarily well-conditioned.

[10] describes an estimator obtained by maximising l_1-penalised likelihood. The resulting estimator has sparse inverses, and is well-conditioned. However, the estimation procedure is computationally expensive, and again, a hyper-parameter is required.

3. Estimating the shrinkage parameter

Recall that for the case where the shrinkage target D consists of the off-diagonal elements of S, the formula (12) gives

$$\lambda = \frac{\sum_{i \neq j} \mathrm{var}\, S_{ij}}{\mathbb{E} \sum_{i \neq j} S_{ij}^2} \tag{14}$$

Neither the numerator nor the denominator of this expression can be obtained directly, and must themselves be estimated from the training data. For this, we follow the procedure of [8], extended for use with GMMs. We replace $\mathbb{E}S_{ij}^2$ by S_{ij}^2 and $\mathrm{var}(S_{ij})$ by its sample variance. An important result from [7] is that this estimator is consistent under a much weaker asymptotic assumption: rather than assuming that the number of samples tends to infinity, it is only necessary to assume that the ratio of the number of parameters to the number of samples is bounded.

We now explain how the sample variance of S_{ij} can be obtained within the context of the EM algorithm. In what follows, we suppress the dependence on the Gaussian m for clarity of notation. We define

$$w_{ij}(t) = (x_i(t) - \hat{\mu}_i)(x_j(t) - \hat{\mu}_j) \tag{15}$$

The sample mean of this is given by

$$\bar{w}_{ij} = \frac{1}{\beta} \sum_t \gamma(t) w_{ij}(t) \tag{16}$$

And the sample estimate of $\mathrm{var}(w_{ij})$ is given by

$$\widehat{\mathrm{var}}(w_{ij}) = \frac{1}{\beta} \sum_t \gamma(t)(w_{ij}(t) - \bar{w}_{ij})^2 \tag{17}$$

$$= \frac{\sum_t \gamma(t) w_{ij}(t)}{\beta} - \bar{w}_{ij}^2 \tag{18}$$

Now since

$$S_{ij} = \frac{1}{\beta} \sum_t \gamma(t) w_{ij}(t) \tag{19}$$

then treating the $w_{ij}(t)$ as IID random variables, we have

$$\widehat{\mathrm{var}} S_{ij} = \frac{\sum_t \gamma(t)^2}{\beta^2} \widehat{\mathrm{var}} w_{ij} \tag{20}$$

This formula is clearly analogous to that derived in [8] when the true class of each sample is observed. As an aside, it should be noted that the variance estimators used here are slightly biased. This could be remedied by applying a correction factor at each stage, given by

$$\frac{\beta^2}{\beta^2 - \sum_t \gamma(t)^2} \tag{21}$$

However, we found that this correction makes little difference in practice.

The shrinkage estimate is obtained at each iteration of the EM algorithm. It does not affect the algorithm's convergence properties. It can be seen from (20) that the computation of λ requires two additional sets of statistics to be accumulated, namely the sums of w_{ij}^2 and γ^2. The additional computational cost is small compared to that already incurred computing of $\gamma(t)$ for each frame.

4. Experiments

To evaluate the potential benefits for ASR systems of using the shrinkage estimator over the standard full covariance matrix estimator, we carried out phone recognition experiments on the TIMIT corpus. Since our particular interest lies in the case when the amount of training data is small, we conducted experiments where the amount of training data is artificially reduced. Data was removed from the full training set on a phone-by-phone basis, using the high-accuracy hand-alignments available for the corpus. We removed instances of each phone at random across the whole corpus, but in such a way as to ensure that the overall distribution of phones remained constant. In the smallest case, data consisted of just 10% of the full training set. This method of reducing the data meant that the HMMs could not be trained using embedded re-estimation: however, use of this technique is not essential for systems trained on TIMIT.

The system used for the experiments was a standard monophone HMM system 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. Following previous experiments, each full-covariance Gaussian mixture component was initialised from a diagonal-covariance system with the same number of Gaussians (and trained on the same data). Results were obtained on the standard reduced test set of 192 utterances, collapsing the labels to the usual 39-phone set. A bigram language model was used for decoding, with the language model scaling factor and insertion penalty fixed for all experiments.

An important issue to consider is the optimal number of Gaussians to use for each state. The issue is particularly problematic here because the optimal number of Gaussians will be higher when the number of parameters per Gaussian is smaller, as in the diagonal or semi-tied covariance cases, and will also vary with the amount of training data available, and the form of estimator used. Previous experiments indicated that performance of diagonal-covariance systems, trained on the full training set, reaches an approximate peak of 69.6% accuracy with 64 Gaussians per state. Full covariance systems reach a peak at a much lower number of Gaussians – usually around 12 per state, depending on the amount of training data. With this in mind, when experimenting with varying amounts of training data, we fix the number of Gaussians at 12. In the results that follow, the main comparison to be drawn is between the system using the shrinkage covariance estimator, and that using the standard full covariance estimator – these both have the same total number of parameters – diagonal covariance and STC systems with the same number of Gaussians[1] are shown for interest, but these results are not indicative of their best performance. However, [3] found that full-covariance systems can outperform STC systems that have more Gaussians and a greater number of parameters in total.

The results with 12 Gaussians per state are shown in Figure 1 and in Table 1. It can be seen that the system using the shrinkage estimator outperforms all the other systems, for all quantities of training data. The performance of standard full covariance system drops rapidly as the amount of training data is reduced, whilst the shrinkage system maintains its robustness. At 10% data, it continues to outperform the diagonal system. Although results are not given for reasons of space, a similar trend was observed with smaller numbers of Gaussians. The best results are competitive with recently-reported results from

[1] STC transforms are tied at the state level

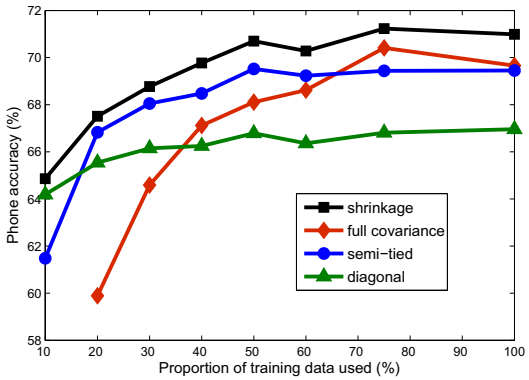

Figure 1: *Phone accuracy of covariance estimators with varying quantities of training data, using 12 Gaussians per state*

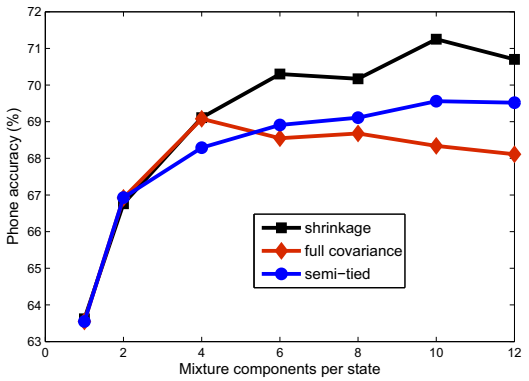

Figure 2: *Phone accuracy of covariance estimators with varying Gaussian mixture components, using 50% of training data*

similar systems on this task [11].

In Figure 2 we illustrate how the performance of the systems vary with the number of Gaussians, in the case when 50% of the training set is used. Again, the shrinkage system outperforms the others in all experiments, whereas the standard full-covariance system lacks robustness at higher number of Gaussians. This is because as the number of Gaussians is increased, the effective amount of data available to train each one is reduced.

5. Discussion

We have demonstrated that the use of the shrinkage estimator can be beneficial for full-covariance ASR systems. This estimator is simple to compute and requires no hyper-parameters to be specified. The method is shown to be particularly beneficial in sparse-data situations.

The sparse-data results shown here are somewhat artificial: to demonstrate the benefits of the technique, it is useful to be able to precisely control the amount of data available. However, there are many ASR applications where the number of parameters is large relative to the amount of training data: for example, when adapting models to new speakers or environments using limited adaptation data; or in systems using an expanded set of acoustic features. We will seek to apply the technique to these tasks in future work.

Table 1: *Phone accuracy results.*

Data (%)	Diag	STC	Full	Shrinkage
10	64.2	61.5	-	64.9
20	65.5	66.8	59.9	67.5
30	66.2	68.1	64.6	68.8
40	66.3	68.5	67.1	69.8
50	66.8	69.5	68.1	70.7
60	66.4	69.2	68.6	70.2
75	66.8	69.4	70.4	71.2
100	67.0	69.5	69.7	71.0

Finally, we note that the methods used do not seek to correct for the invalidity of the model-correctness assumption underpinning MLE. We will therefore seek to adapt the regularisation techniques for use with discriminative estimators, to maintain robustness in limited-data conditions.

6. References

[1] 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.

[2] P. Olsen and R. A. Gopinath, "Modeling inverse covariance matrices by basis expansion," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 37–46, Jan. 2004.

[3] 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, Nov. 2005.

[4] M. Varjokallio and M. Korimo, "Comparison of subspace methods for Gaussian mixture models in speech recognition," in *Proc. Interspeech*, 2007.

[5] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[6] C. Stein, "Inadmissibility of the usual estimator of the mean of a multivariate normal distribution," in *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1956, pp. 197–206.

[7] O. Ledoit and M. Wolf, "A well-conditioned estimator for large covariance matrices," *Journal of Multivariate Analysis*, vol. 88, pp. 365–411, 2004.

[8] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005.

[9] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University Engineering Department, 2003.

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

[11] F. Sha and L. K. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden markov models," in *Proc. ICASSP*, 2007.