On Divergence Based Clustering of Normal Distributions and Its Application to HMM Adaptation

Tor André Myrvoll

Institutt for teleteknikk
NTNU
Trondheim, Norway

Frank K. Soong

Spoken Language Translation Labs
ATR
Kyoto, Japan

Abstract

We present an algorithm for clustering multivariate normal distributions based upon the symmetric, Kullback-Leibler divergence. Optimal mean vector and covariance matrix of the centroid normal distribution are derived and a set of Riccati matrix equations is used to find the optimal covariance matrix. The solutions are found iteratively by alternating the intermediate mean and covariance solutions. Clustering performance of the new algorithm is shown to be superior to that of non-optimal sample mean and covariance solutions. It achieves a lower overall distortion and flatter distributions of pdf samples across clusters. The resultant optimal clusters were further tested on the Wall Street Journal database for adapting HMM parameters in a Structured Maximum A Posterior Linear Regression (SMAPLR) framework. The recognition performance was significantly improved and the word error rate was reduced from 32.6% for a non-optimal centroid (sample mean and covariance) to 27.6% and 27.5% for the diagonal and full covariance matrix cases, respectively.

1. Introduction

The need to cluster multivariate normal distributions is often encountered when working with normal mixture density based Hidden Markov Models (HMMs) in automatic speech recognition (ASR). An indirect way to cluster distributions is frequently used when tied HMMs are constructed using the observation data. Usually the clustering is performed in an agglomerative or divisive hierarchical method with a likelihood based decision rule [1, 2]. A more direct approach that doesn't involve raw observations is to split the parameters of an HMM (i.e., Gaussian kernels) directly into clusters of a hierarchical tree where each node forms its own cluster and the resultant clustered structure is then used for model adaptation algorithms. These algorithms include, e.g., Maximum Likelihood Linear Regression (MLLR)[3], Structural MAP adaptation (SMAP)[4] and Cluster Adaptive Training (CAT)[5]. All these approaches perform well when a logical partitioning of the HMM parameters is carried out, resulting in a richer and more structural mapping of the HMM parameter space.

Most model adaptation algorithms focus only on the mixture component mean vectors of the underlying HMMs. This is due to the fact that the state transition probabilities and mixture weights have little to no effect on the overall recognition performance, and the covariance matrices of the mixture components are numerically unstable to adapt and a robust estimate is difficult to obtain when the adaptation data is scarce. Ideally, we should obtain clusters of mixture components, i.e. pdf distributions, whose mean vectors are structured in such a way that any observed perturbations in a small subset should lead us to infer the adaptation direction and magnitude of the unobserved mean vectors using a simple model, e.g., an affine transformation in the MLLR case. To achieve an effective and meaningful structure "similar" mixture components (i.e., Gaussian kernels) should always be grouped into the same cluster. However, the term "similar" is open for an intuitively appealing and mathematically tractable choice of a distortion measure.

One similarity measure that has been strongly advocated and commonly used e.g., [4, 6] is the divergence measure [7] which measures the "distance" or "distortion" between two given probability density functions, f and g, as

$$d(f,g) = \int f(x) \log \frac{f(x)}{g(x)} dx + \int g(x) \log \frac{g(x)}{f(x)} dx \quad (1)$$

Note that the divergence does not fulfill the triangle inequality, and so it is not a distance or metric, but a legitimate distortion measure as defined in [8]. It is also positive semi-definite. One intuitive interpretation of the symmetric divergence measure is that it is the averaged expected value of the log difference between the two pdf's.

If f and g are multivariate normal distributions, the divergence between them has the closed-form

$$\begin{aligned} & d(f,g) \\ &= \frac{1}{2}\text{trace}\Big\{ (\boldsymbol{\Sigma}_f^{-1} + \boldsymbol{\Sigma}_g^{-1})(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)^T \\ &\qquad + \boldsymbol{\Sigma}_f \boldsymbol{\Sigma}_g^{-1} + \boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_f^{-1} - 2\mathbf{I} \Big\}, \end{aligned} \quad (2)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the corresponding mean vectors and covariance matrices, respectively. In [6] a comparative study was performed on the use of the divergence, Euclidean distance and Bhattacharyya distance as distance measures in constructing a hierarchy of clusters of HMM mixture components. The study shows that the divergence measure gave the best adaptation results. Based upon an earlier paper[9] we will both concentrate on how to find an optimal cluster centroid of multivariate normal distributions using the divergence as the similarity measure, as well investigate its application to HMM adaptation.

2. The Expectation Centroid

The divergence measure has been proposed to measure the similarity between multivariate normal densities for clustering, e.g. [10]. However, no solution has been published, to the authors' best knowledge, on finding the optimal centroid of a cluster of multivariate Gaussian densities with such a choice of distortion measure as the symmetric K-L. divergence. Optimal centroids have

been derived for Bhattacharyya distances [11] and the Kullback-Leibler information measure [12].

The centroid we are interested in is a multivariate normal density that minimizes the total distortions. Formally, a centroid c is defined as,

$$c = \underset{c'}{\operatorname{argmin}} \sum_{n=1}^{N} d(x_n, c'), \tag{3}$$

where N is the number of cluster members, and x_n is the nth cluster member.

In [4] a centroid which we shall refer as the *expectation centroid* is defined as a density whose mean and covariance are the "expected" value of the mean and the covariance of samples,

$$\begin{aligned} \boldsymbol{\mu}_c &= \frac{1}{N}\sum_{n=1}^{N} E[x_n] \\ &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\mu}_n, \end{aligned} \tag{4}$$

$$\begin{aligned} \boldsymbol{\Sigma}_c &= \frac{1}{N}\sum_{n=1}^{N} E[(x_n - \boldsymbol{\mu}_c)(x_n - \boldsymbol{\mu}_c)^T] \\ &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\Sigma}_n + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^T - \boldsymbol{\mu}_c\boldsymbol{\mu}_c^T, \end{aligned} \tag{5}$$

where x_n refers to a random variable distributed according to the nth cluster member. Later experiments show that this centroid has good convergence properties and reasonable clustering performance. However, it is not optimal, i.e., it doesn't minimize the total divergence of a cluster. In the next section we will present an algorithm that find the optimal centroid of a clustered multi-variate normal distribution.

3. The Optimal Centroid

The centroid is also a multivariate normal distribution parameterized by its mean, μ_c, and covariance, Σ_c, which minimizes the overall divergence,

$$\{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\} = \underset{\boldsymbol{\mu}_c', \boldsymbol{\Sigma}_c'}{\operatorname{argmin}} \sum_{n=1}^{N} d(f_{c'}, f_n). \tag{6}$$

Here $d(\cdot,\cdot)$ is the Kullback-Leibler divergence measure, $f_{c'}$ is a multivariate normal distribution with a mean vector $\boldsymbol{\mu}_c'$ and a covariance matrix $\boldsymbol{\Sigma}_c'$, and f_n refers to the nth member in the cluster.

The mean vector of the centroid, $\boldsymbol{\mu}_c$, can be found simply by setting the gradient of the objective function in equation (6) with respect to $\boldsymbol{\mu}_c'$ to zero. This yields the following solution

$$\boldsymbol{\mu}_c = \left[\sum_{n=1}^{N}(\boldsymbol{\Sigma}_n^{-1} + \boldsymbol{\Sigma}_c^{-1})\right]^{-1}\left[\sum_{n=1}^{N}(\boldsymbol{\Sigma}_n^{-1} + \boldsymbol{\Sigma}_c^{-1})\boldsymbol{\mu}_n\right]. \tag{7}$$

Note that this optimal mean vector of the centroid can not be computed directly since it involves Σ_c^{-1}, the inverse of the yet to be determined covariance matrix of the centroid.

The procedure to find the covariance matrix of the centroid is a bit more involved. After some manipulations (see [13] for a complete treatment), it turns out that finding the optimal covariance matrix is equivalent to solving the following Riccati matrix equation,

$$\mathbf{A} + \mathbf{B}\mathbf{X} + \mathbf{X}\mathbf{B}^* - \mathbf{X}\mathbf{C}\mathbf{X} = \mathbf{0}, \tag{8}$$

where

$$\begin{aligned} \mathbf{A} &= \sum_{n=1}^{N}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_c)(\boldsymbol{\mu}_n - \boldsymbol{\mu}_c)^T + \boldsymbol{\Sigma}_n, && (9)\\ \mathbf{B} &= \mathbf{0}, && (10)\\ \mathbf{C} &= \sum_{n=1}^{M}\boldsymbol{\Sigma}_n^{-1}, && (11)\\ \mathbf{X} &= \boldsymbol{\Sigma}_c. && (12) \end{aligned}$$

Equation (8) can be rewritten in a block matrix form as

$$\begin{bmatrix} \mathbf{I} & -\mathbf{X} \end{bmatrix}\begin{bmatrix} \mathbf{B} & \mathbf{A} \\ \mathbf{C} & -\mathbf{B}^* \end{bmatrix}\begin{bmatrix} \mathbf{X} \\ \mathbf{I} \end{bmatrix} = \mathbf{0}. \tag{13}$$

Let us denote the $2d \times 2d$ square block matrix as $\mathbf{M}$. The following theorem is helpful for finding the optimal covariance matrix [14]:

Theorem 3.1 *Assume that* $\mathbf{A}$, $\mathbf{C}$ *are positive semidefinite Hermitian, and let* $\mathbf{v}_1, \ldots, \mathbf{v}_d$ *be the eigenvectors of* $\mathbf{M}$ *corresponding to the eigenvalues* $\lambda_1, \ldots, \lambda_d$. *Further, if* $\mathbf{v}$ *is an eigenvector of* $\mathbf{M}$, *we will write*

$$\mathbf{v} = \begin{bmatrix} \mathbf{u} \\ \mathbf{w} \end{bmatrix},$$

where $\mathbf{u}$ *and* $\mathbf{w}$ *are the upper and lower halves of* $\mathbf{v}$, *respectively. Then, if* $\lambda_1, \ldots, \lambda_d$ *have positive real parts and* $[\mathbf{w}_1, \ldots, \mathbf{w}_d]$ *is nonsingular, the matrix*

$$\mathbf{X} = [\mathbf{u}_1, \ldots, \mathbf{u}_d][\mathbf{w}_1, \ldots, \mathbf{w}_d]^{-1}, \tag{14}$$

is a positive semidefinite solution to equation (13).

For a proof of the existence of d positive eigenvalues, $\lambda_1, \ldots, \lambda_d$, as well as the non-singularity of $[\mathbf{w}_1, \ldots, \mathbf{w}_d]$, we refer to [13]. In the same reference it is also shown that the objective function is convex in both $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$, hence guaranteeing that if a minimum exists, it is global.

In the special case where all the distributions have a diagonal covariance matrix we can constrain the covariance of the centroid to be diagonal, yielding the following simple expressions for the ith elements of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ respectively,

$$\boldsymbol{\mu}_c(i) = \frac{\sum_{n=1}^{N}\left(\boldsymbol{\Sigma}_c^{-1}(i) + \boldsymbol{\Sigma}_n^{-1}(i)\right)\boldsymbol{\mu}_n(i)}{\sum_{k=1}^{N}\left(\boldsymbol{\Sigma}_c^{-1}(i) + \boldsymbol{\Sigma}_k^{-1}(i)\right)}, \tag{15}$$

$$\boldsymbol{\Sigma}_c(i) = \sqrt{\frac{\sum_{n=1}^{N}\boldsymbol{\Sigma}_n(i) + (\boldsymbol{\mu}_c(i) - \boldsymbol{\mu}_n(i))^2}{\sum_{k=1}^{N}\boldsymbol{\Sigma}_k^{-1}(i)}} \tag{16}$$

It should still be noted that the solutions for $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are dependent on each other. No joint solution is readily obtainable, and we compute the $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ iteratively, starting from the expectation centroid.

4. Experimental Results

In this section we present two sets of experimental results. The first result is on the clustering performance where we compare the true centroids, in both forms of full and diagonal covariance matrix, with the expectation centroid defined in equations (4) and (5). In the second experiment we use the true centroid and the expectation centroid to build regression trees and compare the their performance in SMAPLR adaptation [15].

The multivariate normal distributions used in these experiments are from HMMs trained on the speech data from 84 speakers in the Wall Street Journal Corpus. The model contains 37,786 mixture components – all in the form of multivariate normal distribution with diagonal covariance matrix. The model is constructed using feature vectors of 12 mel-frequency cepstral coefficients together with the normalized log energy plus the Δ- and Δ^2-coefficients, 39 features altogether.

4.1. Clustering Performance

4.1.1. Diagonal covariance matrix data

We now use the clustering algorithm with the divergence measure to group HMM mixture components having diagonal covariance matrices into 10 clusters. To account for the fact that the initial conditions change the final clustering performance, we repeated the experiment five times using different initializations while the same initializations were used for all the three centroid finding procedures.

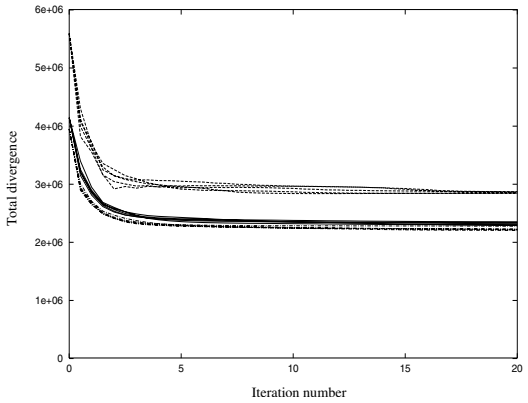

Figure 1: Convergence performance of three different centroids. The dashed lines at the top corresponds to the expectation centroids, the drawn lines in the middle to the optimal diagonally constrained centroids, and the dash-dotted lines, which give the lowest overall distortion, to the optimal centroids with full covariances.

The results can be seen in figure 1. As expected, the two optimal centroids clearly outperform the expectation centroid in terms of the total divergence, with the full covariance centroid being slightly better than the diagonal covariance centroid. Note the expectation based centroid exhibits a non-monotonic decrease of distortion, which is to be expected as the computation of a new set of centroids is not guaranteed to lower the total distortions, as opposed to the true centroids.

The true centroids also yield more evenly distributed clusters of the mixture components, which can be seen from the example in figure 2. A more formal evaluation of the flatness is given in table 1, where the Shannon entropy is used to measure the flatness of the cluster distribution for the five different experiments.

4.1.2. Full covariance matrix data

To see if the approach works as well for pdfs having full covariance matrices, we generated a set of multivariate normal distributions with full covariance matrices using the same WSJ speech training data as in the previous section. An HMM was used to obtain a state level alignment of the WSJ training data.

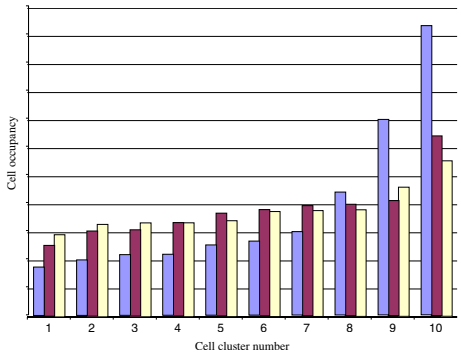

Figure 2: Distribution of mixture components across clusters. Left bar corresponds to expectation centroid, middle to diagonally constrained centroid and right to full covariance centroid.

Trial	Expectation	Diagonal	Full
1	3.0363	3.2766	3.2972
2	2.9978	3.3028	3.2964
3	3.0114	3.2782	3.2332
4	3.0136	3.2599	3.2791
5	3.0002	3.2724	3.2768
Avg.	3.0119	3.2780	3.2765

Table 1: The flatness of the mixture components distribution across clusters, as measured by the entropy of the normalized bin counts. Results of five different trials are presented as well as their average.

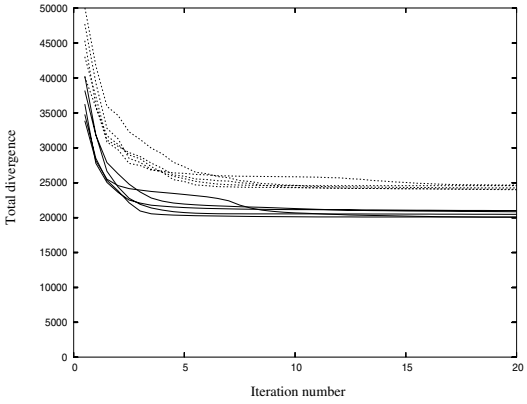

Figure 3: Convergence plots for the full covariance experiment. Five different initializations were used for each centroid. The dashed lines represents the expectation centroid, whereas the continuous lines represent the optimal centroid.

The frames corresponding to each state were then used to estimate 3,448 multivariate normal densities with full covariance matrices.

The clustering experiment was repeated on the new data using the expectation centroid and the optimal centroid. The results are shown in figure 3. As we see the optimal centroid gives us significantly lower total distortions than the expectation centroid. Also, it should be mentioned that the expectation centroid had some initialization problems in this case as many clusters ended up containing only a few or no elements.

4.2. SMAPLR Adaptation

The clustering performance results are definitely encouraging, but it is yet to be confirmed that the optimal centroids can improve the recognition performance, say, in a model adaptation task. The three centroids from the previous clustering experiments were used to build three sets of hierarchical clusterings of HMM mixture components in a form of regression tree. These trees were then used with the SMAPLR approach on the Wall Street Journal Spoke III task, consisting of ten non-native speakers of American English. The original, speaker independent model is the same as the one that we used for our clustering experiments in the previous section. The un-adapted model gives yields baseline performance of 29.2% word error rate (WER). Only one adaptation utterance was used to adapt the model in this adaptation experiment. Earlier experiment has shown that we need to be very conservative in adapting the HMM parameters when data is scarce, as the performance can degrade significantly when a poor mapping is obtained. Here we try to be more ambitious such that we do not end up using only one global transformation, something that would render the experiment meaningless. Just like the clustering experiments we repeated the tree building procedure with five different initializations. The results are presented in table 2.

Trial	Expectation	Diagonal	Full
1	31.6	27.1	27.9
2	33.4	27.5	27.2
3	34.7	27.5	27.9
4	31.7	27.8	26.7
5	31.7	27.9	27.7
Avg.	32.6	27.6	27.5

Table 2: The word error rate (WER) on the Wall Street Journal Spoke III task for three different centroids based hierarchical clustering. Results for five separate experiments are presented, as well as their averages.

The results clearly indicates that the optimal centroids based hierarchical tree clusterings are better suited for adaptation than the trees built using the expectation centroids. While it may still be difficult to pinpoint the exact reason why the adaptation performance is much inferior when using the expectation centroid based trees, as opposed to the trees built using the optimal centroids, we feel rather confident that a more consistent minimization of the distortion and the resultant optimal centroids in the information theoretic sense, lead to better regression trees for adaptation.

5. Conclusion

In this work we present a novel clustering algorithm for finding the optimal centroids of multivariate normal distributions using the Kullback-Leibler divergence measure. It is shown that the clusterings obtained using the optimal centroid yield significantly lower overall distortion than the centroid based on the sample mean and the sample covariance. Also the multivariate normal distributions are more evenly distributed across resultant clusters. The optimal centroids were further used to construct hierarchical regression trees and tested for adapting HMM parameters. The adaptation result shows a clear improvement in WER when compared with the non-optimal centroids.

6. References

[1] A. Kannan, M. Ostendorf, and J. R. Rohlicek, "Maximum likelihood clustering of Gaussians for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, Jul. 1994.

[2] J. J. Odell, P. C. Woodland, and S. J. Young, "Tree-based state clustering for large vocabulary speech recognition," in *Int. Symp. on Speech, Image Proc. and Neural Networks*, Hong Kong, Apr. 1994, pp. 690–693.

[3] C. J. Leggetter and P. C. Woodland, "Speaker adaptation of continuous density HMMs using multivariate linear regression," in *Proc. ICSLP*, Yokahama, Japan, Sep. 1994.

[4] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Transactions on Speech and Audio Processing*, 2000.

[5] Mark. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, Jul. 2000.

[6] J.-T. Chien, "Online hierarchical transformation of hidden Markov models for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 656–667, Nov. 1999.

[7] Solomon Kullback, *Information Theory and Statistics*, Dover Publications, 1997.

[8] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[9] T. A. Myrvoll and F. K. Soong, "Optimal clustering of multivariate normal distributions using divergence and its application to hmm adaptation," in *Proc. IEEE ICASSP-03*, Hong Kong, China, April 2003.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, Inc, 2 edition, 2001.

[11] L. Rigazio, B. Tsakam, and J.-C. Junqua, "An optimal Bhattacharyya centroid algorithm for Gaussian clustering with applications in automatic speech recognition," in *Proc. IEEE ICASSP-00*, Istanbul,Turkey, August 2000.

[12] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *EuroSpeech '97*, Rhodes, Greece, Sep. 1997.

[13] Tor André Myrvoll, *Adaptation of Hidden Markov Models using Maximum a Posteriori Linear Regression with Hierarchical Priors*, Ph.D. thesis, Norwegian University of Science and Technology, 2002.

[14] James E. Potter, "Matrix quadratic solutions," *SIAM J. Appl. Math*, vol. 14, no. 3, pp. 496–501, 1966.

[15] Olivier Siohan, Tor André Myrvoll, and Chin-Hui Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech and Language*, vol. 16, no. 1, pp. 5–25, Jan. 2002.