MAXIMUM LIKELIHOOD ESTIMATES FOR EXPONENTIAL TYPE DENSITY FAMILIES

Sankar Basu, Charles A. Micchelli and Peder A. Olsen

IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598

ABSTRACT

We consider a parametric family of density functions of the type $\exp(-|x|^{\frac{\alpha}{2}})$ for modeling acoustic feature vectors used in automatic recognition of speech. The parameter α is a measure of the impulsiveness as well as the nongaussian nature of the data. While previous work has focussed on estimating the mean and the variance of the data here we attempt to estimate the impulsiveness α from the data on a maximum likelihood basis. We show that there is a balance between α and the number of data points N that must be satisfied before maximum likelihood estimation is carried out. Numerical experiments are performed on multidimensional vectors obtained from speech data.

1. INTRODUCTION

In [2] one considers mixtures of densities of the form

$$f(x) = \frac{\rho}{\sqrt{\sigma}} \exp(-h(\gamma \frac{(x-\mu)^2}{\sigma})),$$

where ρ and γ are chosen so that

$$\int f(x)dx = \frac{1}{\sigma}\int (x-\mu)^2 f(x)dx = 1. \qquad (1)$$

Particular attention was given to the choice $h(t) = t^{\alpha/2}$, $t > 0$, $\alpha > 0$. This particular choice of density has been studied in the literature and is known as alpha-stable densities (or α densities for short) as well as power exponential distributions, cf. [8, 9, 10, 11]. As a consequence of encouraging results from using $\alpha < 0.5$ in [2] we became interested in finding the "optimal" value of α. In so doing, this leads us to a discussion of how one would go about finding the "optimal" value, and whether convergence to the correct value of α could be expected for large amounts of data. We ares also interested in using different values of the parameter α for different mixture components when mixture densities are used for modeling purposes.

2. SOME ANALYTICAL CONSIDERATIONS

For simplicity, first consider the one-component mixture case and discuss whether maximization of the log-likelihood function

$$L(\mu, \sigma, \alpha) = \prod_{i=1}^{N} f(x^i; \mu, \sigma, \alpha)$$

yields the correct α as $N \to \infty$. We take $\{x^i\}_{i=1}^N$ to be one-dimensional data. The maximum likelihood tells us that maximizing $L(\mu, \sigma, \alpha)$ yields consistent estimates for μ, σ and α (see [6]).

We shall attempt to verify if this is so for the densities we are considering. As the logarithm of L is strictly increasing, it suffices to maximize

$$\frac{1}{N}\log L(\mu, \sigma, \alpha) = \frac{1}{N}\sum_{i=1}^{N} \log f(x^i; \mu, \sigma, \alpha).$$

As $N \to \infty$, it is reasonable to expect that

$$\frac{1}{N}\sum_{i=1}^{N} \log f(x^i; \mu, \sigma, \alpha)$$

converges almost surely (in probability) to

$$H(\mu, \sigma, \alpha) = E\{\log f(X; \mu, \sigma, \alpha)\},$$

where X is the random variable from which the samples x_i are drawn. To verify that maximizing $E\{\log f(X; \mu, \sigma, \alpha)\}$ is the sane thing to do we assume the random variable X has the probability density $f(X; \tilde{\mu}, \tilde{\sigma}, \tilde{\alpha})$. In this case, the optimal choice for μ, σ, α ought to be $\tilde{\mu}$, $\tilde{\sigma}$ and $\tilde{\alpha}$.

As mentioned before, $\mu = \tilde{\mu}$, $\sigma = \tilde{\sigma}$ and $\alpha = \tilde{\alpha}$ ought to be the global maximum for $H(\mu, \sigma, \alpha)$. However, not knowing the true value of $\tilde{\mu}$, $\tilde{\sigma}$ and $\tilde{\alpha}$ we can, of course, not compute $H(\mu, \sigma, \alpha)$ and are left with maximizing $L(\mu, \sigma, \alpha) = \prod_{i=1}^{N} f(x^i; \mu, \sigma, \alpha)$. This is equivalent to maximizing

$$\frac{1}{N}\log L = \log \rho - \frac{1}{2}\log \sigma - \left(\frac{\gamma}{\sigma}\right)^{\alpha/2} \frac{1}{N}\sum_{i=1}^{N} |x^i - \mu|^\alpha.$$

One would expect that $\frac{1}{N}\log L(\mu, \sigma, \alpha) \approx H(\mu, \sigma, \alpha)$ and, therefore, that the maximum would occur at $\mu = \tilde{\mu}$, $\sigma = \tilde{\sigma}$ and $\alpha = \tilde{\alpha}$ for large values of N. At this point we state the following result without proof.

Lemma 1 Let $\mathcal{L}_\sigma(\mu, \alpha) = \max_{\sigma \geq 0} \frac{1}{N}\log L(\mu, \sigma, \alpha)$. Then $\mathcal{L}_\sigma(\mu, \alpha) \to \infty$ as $\alpha \to 0$

The last lemma shows that $\max_{\mu,\sigma} \frac{1}{N}\log L(\mu, \sigma, \alpha)$ goes to ∞ as $\alpha \to 0$. However, it can be shown that this is not the case for $H(\mu, \sigma, \alpha)$. Two salient features of this fact are as follows. First, one must, therefore, take care that the value of α does not become too small when seeking a maximum of $\frac{1}{N}\log L$. Next, strictly speaking maximizing $\mathcal{L}_\sigma(\mu, \alpha)$ and $H(\mu, \sigma, \alpha)$ do not satisfy identical objectives.

The authors were supported in part by DARPA Contract No. MDA972-97-C-0012

3. EM TYPE STRATEGIES

For applications to speech recognition, the data is multidimensional and of a nature so complex as not to be accurately described in a single α-density. In [2] various mixtures of multidimensional α-densities were introduced and successfully applied to speech recognition. However, the value of α was fixed a priori and left constant over the mixture components. We will discuss how the individual mixture components can have differing values of α and how one goes about finding the optimal choices of α.

Let us describe how mixtures of multidimensional α-densities are constructed. The individual components are given by

$$p(x|\lambda^\ell) = \frac{\rho_d(\alpha^\ell)}{\sqrt{\prod_{i=1}^d \sigma_i^\ell}} e^{-\left(\gamma_d(\alpha^\ell)\sum_{i=1}^d \frac{(x_i-\mu_i^\ell)^2}{\sigma_i^\ell}\right)^{\alpha^\ell/2}} \qquad (2)$$

where

$$\rho_d(\alpha) = \frac{\alpha}{2}\frac{,(\frac{d}{2})}{(d\pi)^{\frac{d}{2}}}\frac{,(\frac{d+2}{\alpha})^{\frac{d}{2}}}{,(\frac{d}{\alpha})^{\frac{d}{2}+1}}; \text{ and } \gamma_d(\alpha) = \frac{,(\frac{d+2}{\alpha})}{d,(\frac{d}{\alpha})}$$

and λ^ℓ denotes the collection of parameters α^ℓ, μ^ℓ and σ^ℓ, where $\ell = 1, \ldots, m$. The mixture density is now given by

$$P(x|\Lambda, \omega) = \sum_{\ell=1}^m \omega^\ell p(x|\lambda^\ell).$$

The log-likelihood of a data set $\{x^k\}_{k=1}^N$ is, thus, given as

$$\log L = \sum_{k=1}^N \log\left(\sum_{\ell=1}^m \omega^\ell p(x|\lambda^\ell)\right).$$

We are ultimately interested in maximizing $\log L$. A desirable property of an iteration scheme would, therefore, be to increase the value of $\log L$. We denote old parameters by 'hatted' quantities and and mimic the EM philosophy as expounded in [5]. We have

$$\begin{aligned}\log L - \log \hat{L} &= \sum_{k=1}^N \log\left\{\frac{\sum_{\ell=1}^m \omega^\ell p(x^k|\lambda^\ell)}{\sum_{j=1}^m \hat{\omega}^j p(x^k|\hat{\lambda}^j)} \times \frac{\hat{\omega}^\ell p(x^k|\hat{\lambda}^\ell)}{\hat{\omega}^\ell p(x^k|\hat{\lambda}^\ell)}\right\} \\ &\geq \sum_{k=1}^N \sum_{\ell=1}^m \frac{\hat{\omega}^\ell p(x^k|\hat{\lambda}^\ell)}{\sum_{j=1}^m \hat{\omega}^j p(x^k|\hat{\lambda}^j)} \times \\ &\quad \log\left\{\frac{\omega^\ell p(x^k|\lambda^\ell)}{\hat{\omega}^j p(x^k|\hat{\lambda}^j)}\right\}\end{aligned}$$

where the well known Jensen's inequality [1] arising from the concavity of the logarithmic function has been used in the last step with $b_\ell = \hat{\omega}^\ell p(x^k|\hat{\lambda}^\ell)\left(\sum_{j=1}^m \hat{\omega}^j p(x^k|\hat{\lambda}^j)\right)^{-1}$ for $\ell = 1, \ldots, m$. We regroup the last equation into three types of terms.

$$\log L - \log \hat{L} \leq A + B + C$$

where

$$A(\omega, \hat{\omega}, \hat{\Lambda}) = \sum_{\ell=1}^m A_\ell \log(\omega^\ell)$$

$$B(\Lambda, \hat{\omega}, \hat{\Lambda}) = \sum_{k=1}^N \sum_{\ell=1}^m A_{\ell k} \log p(x^k|\lambda^\ell)$$

$$C(\hat{\omega}, \hat{\Lambda}) = \sum_{k=1}^N \sum_{\ell=1}^m A_{\ell k} \log(\hat{\omega}^\ell p(x^k|\hat{\lambda}^\ell))$$

and

$$A_{\ell k} = \frac{\hat{\omega}^\ell p(x^k|\hat{\lambda}^\ell)}{\sum_{j=1}^m \hat{\omega}^j p(x^k|\hat{\lambda}^j)} \text{ and } A_\ell = \sum_{k=1}^N A_{\ell k}.$$

Note that the term C only depends on old parameters and A depends only on ω^ℓ, $\ell = 1, \ldots, m$ and old parameters, whereas B depends on Λ and old parameters. It is imporatnt to note for our purposes that only B depends on the new values of α, namely, $\hat{\alpha}$ to be updated. Clearly, $(\log L - \log \hat{L}) = 0$ when the old parameters and the new parameters are equal. Maximizing $(\log L - \log \hat{L})$ for a particular parameter while the others are fixed guarantees that the log-likelihood does not decrease. This can be done explicity for ω^ℓ, $\ell = 1, \ldots, m$ subject to the constraint $\sum_{\ell=1}^m \omega^\ell = 1$.

Using the method of Lagrange multiplier we arrive at the following equation

$$A_\ell - \Theta\omega^\ell = 0, \quad \ell = 1, \ldots, m,$$

where the parameter Θ is the Lagrange multiplier. Solving for Θ we get $\Theta = \sum_{\ell=1}^m A_\ell = N$, which yields $\omega^\ell = \frac{1}{N}A_\ell$. This was done in [2]. Similarly, one may try to maximize with respect to μ^ℓ and σ^ℓ, but this cannot be done explicitly. The stationary equation is available in [2] and the update formulas for iteratively computing μ_i^ℓ and σ_i^ℓ are exactly analogous to those in [2] except that now the parameter α in considering the ℓ-th mixture component has to be replaced by α^ℓ for all $\ell = 1, \ldots, m$. It remains to construct update formulas for α^ℓ for $\ell = 1, \ldots, m$. We have

$$\begin{aligned}\log p(x|\lambda^\ell) &= \frac{1}{2}\left(\sum_{i=1}^d \log \sigma_i^\ell\right) + \log \rho_d(\alpha^\ell) \qquad (3) \\ &\quad - \left(\gamma_d(\alpha^\ell)\sum_{i=1}^d \frac{(x_i-\mu_i^\ell)^2}{\sigma_i^\ell}\right)^{\alpha^\ell/2}\end{aligned}$$

which makes it possible to separate the α^ℓ variables.

$$B(\Lambda, \hat{\omega}, \hat{\Lambda}) = \sum_{\ell=1}^m B^\ell(\Lambda, \hat{\omega}, \hat{\Lambda})$$

where

$$\begin{aligned}B^\ell(\Lambda, \hat{\omega}, \hat{\Lambda}) &= \sum_{k=1}^N A_{\ell k}\left(\frac{1}{2}\left(\sum_{i=1}^d \log \sigma_i^\ell\right) + \log \rho_d(\alpha^\ell)\right. \\ &\quad \left. - \left(\gamma_d(\alpha^\ell)\right)^{\alpha^\ell/2}\left(\sum_{i=1}^d \frac{(x_i^k-\mu_i^\ell)^2}{\sigma_i^\ell}\right)^{\alpha^\ell/2}\right)\end{aligned}$$

Since only B depends on new values of the parameter α, to maximize $(\log L - \log \hat{L})$ with respect to α^ℓ, it suffices to maximize $B^\ell(\Lambda, \hat{\omega}, \hat{\Lambda})$ with respect to α^ℓ. This can be done

numerically. However, we decided to maximize $B^{\ell}(\Lambda, \hat{\omega}, \hat{\Lambda})$ by brute force. Note that this can be done without incurring much computational cost. Assuming that we wish to compute $B^{\ell}(\Lambda, \hat{\omega}, \hat{\Lambda})$ for $\alpha_{\ell} = \alpha_{\min}, \alpha_{\min} + \Delta_{\alpha}, \ldots, \alpha_{\min} + N_{\alpha}\Delta_{\alpha} = \alpha_{\max}$ we note that the greatest computational cost is in computing $S_{k\ell} = \sum_{i=1}^{d} \frac{(x_i^k - \mu_i^{\ell})^2}{\sigma_i^{\ell}}$ for $k = 1, 2, \ldots, N$ and $\ell = 1, 2, \ldots, m$. Once $(S_{k\ell})^{\Delta_{\alpha}}$ and $(S_{k\ell})^{\alpha_{\min}}$ have been computed the quantities $(S_{k\ell})^{\alpha_{\min} + j\Delta_{\alpha}}$ can easily be computed from the corresponding value for $(j-1)$ by one single multiplication. In any case, as we are maximizing $(\log L - \log \hat{L})$ over a discrete set of α's, that contain the previous value of α, we are guaranteed that the log-likelihood is nondecreasing.

4. SPEECH RECOGNITION EXPERIMENTS

Digitized speech sampled at a rate of 16 Khz is considered. A frame consists of a segment of speech of duration 25 msec, and produces an 39 dimensional acoustic cepstral vector via the following process, which is standard in speech recognition literature. Frames are advanced every 10 msec to obtain succeeding acoustic vectors.

First, magnitudes of discrete Fourier transform of samples of speech data in a frame are considered in a logarithmically warped frequency scale. Next, these amplitude values themselves are transformed to a logarithmic scale, and subsequently, a rotation in the form of discrete cosine transform is applied. The first 13 components of the resulting vector are retained. First and the second order differences of the sequence of vectors so obtained are then appended to the original vector to obtain the 39 dimensional cesptral acoustic vector.

As in supervised learning tasks, we assume that these vectors are labeled according to the basic sounds they correspond to. In fact, the set of 46 phonemes are subdivided into a set of 126 different variants each corresponding to a 'state' in the hidden Markov model used for recognition purposes. They are further subdivided into more elemental sounds called allophones or leaves by using the method of decision trees depending on the context in which they occur, (see, e.g., [3, 4, 7] for more details).

Two measurable quantities to evaluate our technique for optimizing α are average log-likelihood and performance of the speech recognizer. We deal with the former first. The data used is from a specific leaf (to be specific, leaf no. 513). We computed the log-likelihood after each iteration with and without using the update formula for α. We found that the likelihood gain was considerable $\alpha = 2$. The conclusion by examining the graphs for the log-likelihood is that the update formula for α gives consistent improvement in log-likelihood. See Figure 1.

As the ultimate objective in speech recognition is to discriminate different sounds, we decided to evaluate the discriminatory power of our density estimates. To this end, we evaluate the densities for all allophones (there are approximately 3500 of them) and compare the density of the "correct" allophone with all the others. If the correct allophone yields higher likelihood value than all the other allophone, we indeed achieve our goal. We produce frequencies for the correct leaf to be among the top 1, 10, 100 and 1000

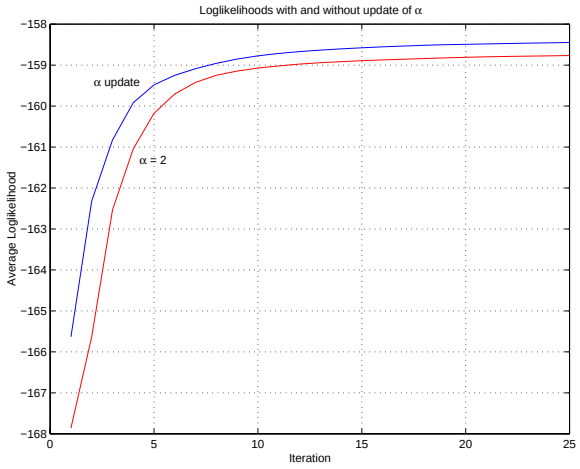

Figure 1: Comparison of α update vs non-update for $\alpha = 2$

highest densities. These numbers are displayed in Table 1. As can be seen, the discriminatory power of the scheme with updated α's is significantly better than without updating α.

Preliminary recognition experiments were carried out on the broadcast transcription task [12] by allowing different mixture components to have different values of the parameter α as compared with the fixed values $\alpha = 1$ and $\alpha = 2$. The results of this experiment for different acoustic conditions are tabulated Table 2. In this table, the various acoustic conditions can be described as: prepared speech (F0), spontaneous speech (F1), low fidelity (telephone) speech (F2), speech with background music (F4), and non-native speakers (F5).

The distribution of α, which was constrained to lie between 0.10 and 2.0 for approximately 121,000 mixture components is shown in Figure 2. Note that preferred values of α tends to be less that 1.0, confirming on a systematic basis that nongaussian mixture components are preferred.

5. CONCLUSION

We have addressed the issue of finding the optimal value of α in densities of the type (1) for speech data. Furthermore, the strategy of allowing different mixture components to have different α-values were also examined in the context of LVCSR. The results indicate a clear departure from gaussian mixture modeling.

6. REFERENCES

[1] M. Abramowitz and I. Stegan, Handbook of Mathematical Functions, Dover Publications, New York, Ninth Dover printing, 1972.

[2] S. Basu and C.A. Micchelli, Parametric density estimation for the classification of acoustic feature vectors in speech recognition, in Nonlinear Modeling: Advanced Black-Box Techniques (Eds. J. A. K. Suykens and J. Vandewalle), pp. 87-118, Kluwer Academic Publishers, Boston 1998.

Table 1: Leaf discrimination for initial value $\alpha = 1$ for no update of α vs update of α. Columns with headings "1", "10", "100" and "1000" contain the number of vectors for which the correct leaf was among the 1, 10, 100 and 1000 first leaves. Exactly 100 vectors were sampled for each leaf.

Without update of α					
Leaf	1	10	100	1000	Ave.
0	25	78	96	100	20.5
1	24	70	96	100	23.3
2	56	88	98	99	23.2
3	28	83	97	100	13.3
4	29	80	96	100	16.2
5	33	79	97	99	29.1
6	20	66	95	100	23.8
7	15	52	88	100	57.1
8	46	78	95	100	32.4
9	38	76	95	100	21.4

With update of α					
Leaf	1	10	100	1000	Ave.
0	32	74	98	100	16.2
1	24	73	97	100	13.3
2	64	92	98	99	22.6
3	36	88	98	100	8.4
4	25	80	97	100	12.9
5	33	75	98	99	33.2
6	30	71	96	100	18.6
7	21	57	90	99	48.0
8	43	82	94	100	26.3
9	43	78	95	100	15.4

Table 2: Prelimary results showing improvements due to using leaf dependent α as opposed to fixed α

Experiment	All	F0	F1	F2	F3	F4	F5	FX
$\alpha = 2.0$	26.1	11.8	22.9	32.1	27.9	27.7	23.1	43.9
$\alpha = 1.0$	25.5	11.5	23.0	31.3	28.1	27.6	21.6	41.1
Variable α	25.4	11.9	22.6	31.3	29.0	26.5	21.8	41.1

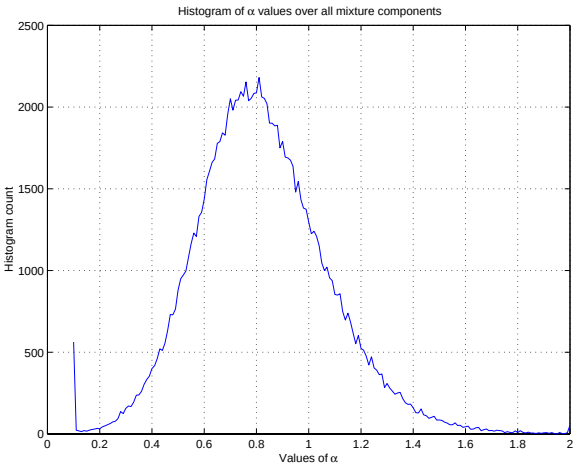

Figure 2: Histogram showing the distribution of α values among different mixture components. Note the central tendency at values near but slightly smaller than 1.0

[3] Frederick Jelenik, Statistical Methods for Speech Recognition, MIT Press, 1997.

[4] L. R. Bahl, P. V. Desouza, P. S. Gopalkrishnan, M. A. Picheny, Context dependent vector quantization for continuous speech recognition, Proceedings of IEEE Int. Conf. on Acosutics Speech and Signal Processing, pp. 632-635, 1993.

[5] Christopher M. Bishop, Neural Networks for Pattern Recognition, Cambridge University Press, 1997.

[6] R. A. Fisher, Contributions to Mathematical Statistics, John Wiley & Sons New York 1950.

[7] Leo Breiman, Classification and Regression Trees, Wadsworth International, Belmont, California, 1983

[8] E. Gòmez, M. A. Gòmez–Villegas, J. M. Marin, A multivariate generalization of the power exponential family of distributions, Comm. Stat. — Theory Meth. 17(3), pp.589–600, 1998.

[9] Owen Kenny, Douglas Nelson, John Bodenschatz and Heather A. McMonagle, Separation of nonspontaneous and spontaneous speech, Proc. ICASSP, 1998.

[10] E. Fama and R. Roll, Parametric estimates for stable distributions, Journal of the American Statistical Association, vol. 66, pp. 331-338, 1971.

[11] J. H. McCulloch, Simple consistent estimators of stable distribution parameters, Communications in Statistics and Simulation, vol. 15, no. 4, pp. 1109-1136, 1986.

[12] L. Polymenakos, P. Olsen, D. Kanevsky, R. A. Gopinath, P. S. Gopalakrishnan and S. Chen, Transcription of broadcast news - some recent improvements to IBM's LVCSR system, ICASSP '98.