Distribution of the Density of a Gaussian Mixture

Jan Larsen
Intelligent Signal Processing Group
Informatics and Mathematical Modelling
Technical University of Denmark
web: isp.imm.dtu.dk, email: jl@imm.dtu.dk

February 19, 2003

1 Introduction

Consider a K component Gaussian mixture density of a feature vector $\boldsymbol{x}$ of dimension d, is defined as

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} P(k) p(\boldsymbol{x}|k, \boldsymbol{\theta}_k) \tag{1}$$

$$p(\boldsymbol{x}|k, \boldsymbol{\theta}_k) = \frac{1}{\sqrt{|2\pi \boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\right) \tag{2}$$

where the component Gaussians are mixed with proportions $\sum_k P(k) = 1, P(k) \geq 0$, and $\boldsymbol{\theta}_k \equiv \{\boldsymbol{\Sigma}_k, \boldsymbol{\mu}_k\}$ is a parameter vector.

The detection of novelty/outliers or evaluating confidence of $p(\boldsymbol{x})$ can be done via

$$Q(t) = \mathrm{Prob}(\boldsymbol{x} \in \mathcal{R}),\ \mathcal{R} = \{\boldsymbol{x} : p(\boldsymbol{x}|k) < t\} \tag{3}$$

which is the distribution function of the density values [2, 3, 4, 5, 6, 7, 9].

Practically, $Q(t)$ can be computed from a surrogate data set, $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^{N}$ of samples drawn from $p(\boldsymbol{x})$. Rank $t_n = p(\boldsymbol{x}_n)$, $\boldsymbol{x}_n \in \mathcal{D}$ in ascending order, $t_1 \leq t_2 \leq \cdots \leq t_N$, and let $Q(t_n) = n/N$. We then set a low threshold $Q_{\min}$ and find the corresponding $t_{\min}$ as $t_{\min} = \arg\min_t Q(t) \geq Q_{\min}$. Finally, novel events are detected as those having density values less than $t_{\min}$.

The aim of this technical report is to device a approximate analytical method, which avoids the generation of a large surrogate data set.

2 Approximate Analytical expression of $Q(t)$

Consider $L(\boldsymbol{x}) = \log p(\boldsymbol{x})$ as a function of the random variable $\boldsymbol{x}$, and define the associated probability density function, $p_L(t)$, and cumulative distribution $P_L(t) = \mathrm{Prob}(L \leq t) = \int_{-\infty}^{t} p_L(s)\, ds$.

To understand the relation between $P_L(t)$ and $Q(t)$, note that $P_L(t)$ is the distribution of log-$p(\boldsymbol{x})$ density values, whereas $Q(t)$ is the distribution of $p(\boldsymbol{x})$ density values. The novelty detection procedure described above could as well be based on $P_L(t)$.

Consider for all $\ell = \arg\max_k P(k)p(\boldsymbol{x}|k)$ and $\boldsymbol{x}$ that

$$\frac{\sum_{k \neq \ell} P(k)p(\boldsymbol{x}|k)}{P(\ell)p(\boldsymbol{x}|\ell)} \ll 1, \tag{4}$$

which means that the distance between clusters are large compared to cluster widths.

Under this assumption[1] using equations (1), (2)

$$\begin{aligned} \log p(\boldsymbol{x}) &= \log\left(\sum_{k=1}^{K} P(k)p(\boldsymbol{x}|k, \boldsymbol{\theta}_k)\right) \\ &\approx \log P(\ell) - \frac{1}{2}\log|2\pi\boldsymbol{\Sigma}_\ell| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_\ell)^\top \boldsymbol{\Sigma}_\ell^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_\ell) \end{aligned} \tag{5}$$

In order to approach the distribution of $L = \log p(\boldsymbol{x})$, recall that a sample from a Gaussian mixture can be obtained by sampling a cluster k with $P(k)$ then sampling $\boldsymbol{x}$ from the corresponding Gaussian $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \sim \chi^2(d)$. Within a specific cluster, ℓ, then according to equation (5)

$$\begin{aligned} \log p(\boldsymbol{x}) &\sim \log P(\ell) - \frac{1}{2}\log|2\pi\boldsymbol{\Sigma}_\ell| - \frac{1}{2}\chi^2(d) \\ &= C_\ell - \frac{1}{2}\chi^2(d), \end{aligned} \tag{6}$$

where $C_\ell = \log P(\ell) - \frac{1}{2}\log|2\pi\boldsymbol{\Sigma}_\ell|$. In consequence, L is approximately a mixture of biased χ^2 distributions

$$p_L(t) = \sum_{k=1}^{K} P(k)p_L(t|k), \tag{7}$$

[1] $\log(a+b) = \log a + \log(1 + b/a) \approx \log a + O(b/a)$.

where $p_L(t|k) \sim C_k - \frac{1}{2}\chi^2(d)$. That is,

$$\begin{aligned}
P_L(t) &= \mathrm{Prob}(L \leq t) = \sum_k P(k)\mathrm{Prob}\left(C_k - \frac{1}{2}\chi^2(d) \leq t\right) \\
&= \sum_k P(k)\mathrm{Prob}\left(\chi^2(d) \geq 2(C_k - t)\right) \\
&= \sum_k P(k)\left(1 - \mathrm{Prob}\left(\chi^2(d) < 2(C_k - t)\right)\right) \\
&= 1 - \sum_k P(k)P_\chi(2(C_k - t); d)), \qquad (8)
\end{aligned}$$

where $P_\chi(t;n)$ is the cumulative distribution of a χ^2-variable with n degrees of freedom given by [1, Ch. 26.4]

$$P_\chi(t;n) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})}\int_0^t s^{\frac{n}{2}-1}e^{-\frac{s}{2}}\,ds,\ t \geq 0 \qquad (9)$$

which essentially is a scaled incomplete gamma function [1, Ch. 6.5.1]. When $t \leq 0$ then $P_\chi(t;n) = 0$, this means that $C_k > t$ should in the terms of equation (8) to give non-zero contributions.

2.1 Example

Consider a $d = 1$ mixture of Gaussian distribution with $K = 2$, $\mu_1 = 0$, $\mu_2 = s$, $\sigma_1 = \sigma_2 = 1$. The evaluation of the approximation [8] is shown in figure 1.

References

[1] M. Abramowitz and I. A. Stegun: *Handbook of Mathematical Functions*, Dover Publications Inc., 1970.

[2] Baker L.D., Hofmann T., Maccallum A.K., and Yang Y. (1999) A Hierarchical Probabilistic Model for Novelty Detection in Text, *CMU techincal report*, `http://www.cs.cmu.edu/People/mccallum/papers/tdt-nips99s.ps.gz`

[3] Basseville M., and Nikiforov I.V. (1993) *Detection of Abrupt Changes: Theory and Application*, Prentice-Hall.

[4] Bishop C.M. (1994) Novelty Detection and Neural Network Validation, *IEE Proceedings - Vision Image and Signal Processing*, vol. 141, no. 4, pp. 217–222.

[5] Box G.E.P., and Tiao G.C. 1992 *Bayesian Inference in Statistical Analysis*, John Wiley & Sons.

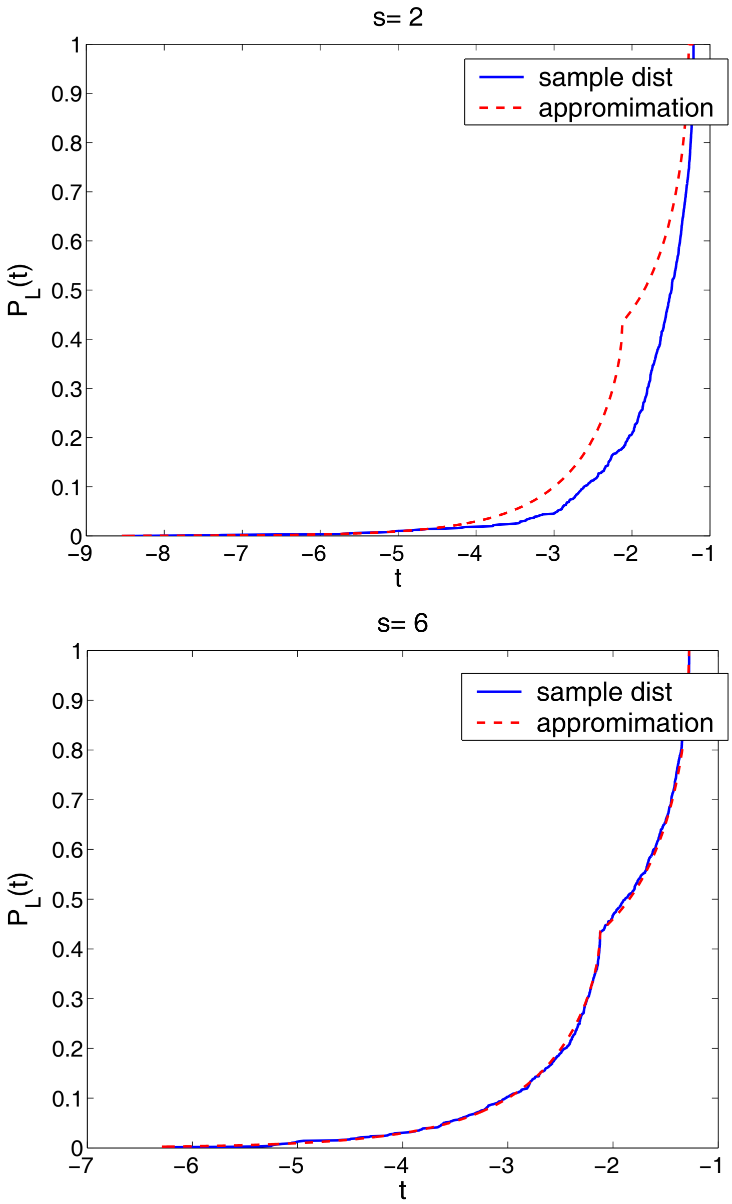

Figure 1: Evaluation of the approximation of $P_L(t)$ for a one dimensional two component Gaussian mixture mode. s is the distance between the components.

[6] J. Larsen, L.K. Hansen, A. Szymkowiak, T. Christiansen and T. Kolenda: "Webmining: Learning from the World Wide Web," invited contribution for *Proceedings of Nonlinear Methods and Data Mining 2000*, Rome, Italy, Sept. 25–26, 2000, pp. 106–125

[7] J. Larsen, L.K. Hansen, A. Szymkowiak, T. Christiansen and T. Kolenda: "Webmining: Learning from the World Wide Web", in special issue of *Computational Statistics and Data*

Analysis, vol. 38, pp. 517–532, 2002.

[8] J. Larsen: Matlab function `qfcttest.m`, Informatics and Mathematical Modelling, Tehcnival University of Denmark, February 2003. Zip-file availble from `http://www.imm.dtu.dk/pubdb/p.php?1755`

[9] Nairac A., Corbett-Clark T., Ripley R., Townsend N., and Tarassenko L. (1997) Choosing An Appropriate Model for Novelty Detection, *IEE 5th Int. Conf. on Artificial Neural Networks*, pp. 117–122.