In: Neural Computation 12(1), pp. 141-152 (January 2000)
URL: http://www.dcs.shef.ac.uk/~miguel/papers/mix_bernoulli.html

Practical identifiability of finite mixtures of multivariate Bernoulli distributions

Miguel Á. Carreira-Perpiñán Steve Renals
Dept. of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
{M.Carreira,S.Renals}@dcs.shef.ac.uk

June 4, 1998

Abstract

The class of finite mixtures of multivariate Bernoulli distributions is known to be nonidentifiable, i.e., different values of the mixture parameters can correspond to exactly the same probability distribution. In principle, this would mean that sample estimates using this model would give rise to different interpretations. We give empirical support to the fact that estimation of this class of mixtures can still produce meaningful results in practice, thus lessening the importance of the identifiability problem. We also show that the EM algorithm is guaranteed to converge to a proper maximum likelihood estimate, owing to a property of the log-likelihood surface. Experiments with synthetic data sets show that an original generating distribution can be estimated from a sample. Experiments with an electropalatography (EPG) data set show important structure in the data.

1 Introduction

Finite mixtures of multivariate Bernoulli distributions have been extensively used in diverse fields (such as bacterial taxonomy) to model a population of binary, multivariate measurements in terms of a few latent classes; see (Everitt and Hand, 1981; Gyllenberg et al., 1994) and references therein. However, it has been recently proven that this class of mixture models is nonidentifiable (Gyllenberg et al., 1994), which potentially undermines the interpretation of sample estimates, since the same sample could be equally attributed to a number of different estimates.

We consider a finite mixture distribution (Everitt and Hand, 1981) defined on the D-dimensional binary space $\{0,1\}^D$:

$$p(\mathbf{t}) = \sum_{m=1}^{M} \pi_m p(\mathbf{t}|m)$$

where M is the number of components, $\mathbf{t} = (t_1 \dots t_D)^T$ is a binary D-dimensional vector, the $\pi_m = p(m)$ are the mixture proportions and $p(\mathbf{t}|m)$ is a multivariate Bernoulli distribution with parameters (called *prototypes*) $\mathbf{p}_m = (p_{m1}, \dots, p_{mD})^T$, that is,

$$p(\mathbf{t}|m) = \prod_{d=1}^{D} p_{md}^{t_d} (1-p_{md})^{1-t_d}.$$

These parameters must obey the following constraints:

$$\sum_{m=1}^{M} \pi_m = 1; \qquad \pi_m \in (0,1) \quad \forall m = 1,\dots,M; \qquad p_{md} \in [0,1] \quad \forall m = 1,\dots,M,\ d = 1,\dots,D.$$

It has been proven (Gyllenberg et al., 1994) that this class of mixtures is nontrivially nonidentifiable[1] for all dimensions D. This means that there are many combinations of values of the parameter tuple $\boldsymbol{\Theta} = \{M,\ \{\pi_m, \mathbf{p}_m\}_{m=1}^{M}\}$ that produce an identical distribution $p(\mathbf{t})$: there exist at least two tuples $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}'$ for

[1] As opposed to trivial nonidentifiability, which is given by permutations of the mixture components or by coincident component distributions $p(\mathbf{t}|m)$ for several components.

which $p(\mathbf{t}|\boldsymbol{\Theta}) = p(\mathbf{t}|\boldsymbol{\Theta}') \ \forall \mathbf{t} \in \{0,1\}^D$. For example, it can be readily verified that the four mixtures given by the following parameter tuples represent the same distribution (here $D = 3$):

$$
\begin{array}{ll}
\boldsymbol{\Theta}: & \big\{M=1, \quad \big\{\pi_1 = 1,\ \mathbf{p}_1 = (\tfrac{1}{2}\ \tfrac{1}{2}\ \tfrac{1}{2})^T\big\}\big\} \\
\boldsymbol{\Theta}': & \big\{M=2, \quad \big\{\pi_1 = \tfrac{1}{2},\ \mathbf{p}_1 = (\tfrac{1}{2}\ 0\ \tfrac{1}{2})^T\big\}, \quad \big\{\pi_2 = \tfrac{1}{2},\ \mathbf{p}_2 = (\tfrac{1}{2}\ 1\ \tfrac{1}{2})^T\big\}\big\} \\
\boldsymbol{\Theta}'': & \big\{M=2, \quad \big\{\pi_1 = \tfrac{1}{4},\ \mathbf{p}_1 = (\tfrac{1}{2}\ 0\ \tfrac{1}{2})^T\big\}, \quad \big\{\pi_2 = \tfrac{3}{4},\ \mathbf{p}_2 = (\tfrac{1}{2}\ \tfrac{2}{3}\ \tfrac{1}{2})^T\big\}\big\} \\
\boldsymbol{\Theta}''': & \big\{M=2, \quad \big\{\pi_1 = \tfrac{1}{4},\ \mathbf{p}_1 = (1\ \tfrac{1}{2}\ \tfrac{1}{2})^T\big\}, \quad \big\{\pi_2 = \tfrac{3}{4},\ \mathbf{p}_2 = (\tfrac{1}{3}\ \tfrac{1}{2}\ \tfrac{1}{2})^T\big\}\big\}\ .
\end{array}
$$

However, it does not mean that for every parameter tuple $\boldsymbol{\Theta}$ there must exist at least one different $\boldsymbol{\Theta}'$ representing the same distribution. Identifiability is a property of the class of mixtures, rather than of a particular parameter tuple.

Hence, in principle there are many tuples in parameter space that are completely equivalent but that would give rise to different interpretations. This may seem an insurmountable difficulty for parameter estimation, but our practical studies have produced promising results: we will show that, given a sample from a mixture of multivariate Bernoulli distributions, maximum likelihood estimates of the parameters (obtained by an EM algorithm) can be still interpretable.

Before giving some experimental results to support this claim, we give some properties of the log-likelihood surface of mixtures of multivariate Bernoulli distributions and introduce an EM algorithm for them.

2 Maximum likelihood parameter estimation

We assume a fixed number of components M. Maximum likelihood estimation can be achieved by an EM algorithm. Define $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)^T$ and $\mathbf{P} = (\mathbf{p}_1, \dots, \mathbf{p}_M)$. The log-likelihood of the parameters $\{\boldsymbol{\pi}, \mathbf{P}\}$ given a sample $\{\mathbf{t}_n\}_{n=1}^N$ is

$$l(\boldsymbol{\pi}, \mathbf{P}) = \sum_{n=1}^N \log p(\mathbf{t}_n; \boldsymbol{\pi}, \mathbf{P}) = \sum_{n=1}^N \log \left(\sum_{m=1}^M \pi_m \prod_{d=1}^D p_{md}^{t_{nd}} (1-p_{md})^{1-t_{nd}} \right) \tag{1}$$

and its gradient is easily seen to be

$$\frac{\partial l}{\partial \pi_m} = \frac{1}{\pi_m} \sum_{n=1}^N p(m|\mathbf{t}_n; \boldsymbol{\pi}, \mathbf{P}) - N \qquad m = 1, \dots, M \tag{2}$$

$$\frac{\partial l}{\partial p_{md}} = \frac{1}{p_{md}(1-p_{md})} \sum_{n=1}^N p(m|\mathbf{t}_n; \boldsymbol{\pi}, \mathbf{P})(t_{nd} - p_{md}) \qquad m = 1, \dots, M \quad d = 1, \dots, D \tag{3}$$

where

$$p(m|\mathbf{t}_n; \boldsymbol{\pi}, \mathbf{P}) = \frac{p(\mathbf{t}_n|m; \boldsymbol{\pi}, \mathbf{P}) p(m)}{\sum_{m'=1}^M p(\mathbf{t}_n|m'; \boldsymbol{\pi}, \mathbf{P}) p(m')} = \frac{\pi_m \prod_{d=1}^D p_{md}^{t_{nd}} (1-p_{md})^{1-t_{nd}}}{\sum_{m'=1}^M \pi_{m'} \prod_{d=1}^D p_{m'd}^{t_{nd}} (1-p_{m'd})^{1-t_{nd}}} \tag{4}$$

are the posterior probabilities (or responsibilities) that component m generated data point $\mathbf{t}_n$. The term "$-N$" in eq. (2) results from the constraint $\sum_{m=1}^M \pi_m = 1$ introduced in the log-likelihood via a Lagrange multiplier.

Derivation of the EM algorithm for finite mixtures of multivariate Bernoulli distributions is straightforward and can be found elsewhere (Everitt and Hand, 1981; Wolfe, 1970). We give here the basic equations:

- E step: computation of the responsibilities using equation (4) from the current parameter estimates $\{\boldsymbol{\pi}^{(\tau)}, \mathbf{P}^{(\tau)}\}$ at iteration τ, $p(m|\mathbf{t}_n; \boldsymbol{\pi}^{(\tau)}, \mathbf{P}^{(\tau)})$.
- M step: reestimation of $\{\boldsymbol{\pi}^{(\tau+1)}, \mathbf{P}^{(\tau+1)}\}$:

$$\pi_m^{(\tau+1)} = \frac{1}{N} \sum_{n=1}^N p(m|\mathbf{t}_n; \boldsymbol{\pi}^{(\tau)}, \mathbf{P}^{(\tau)}) \qquad\qquad \mathbf{p}_m^{(\tau+1)} = \frac{1}{N\pi_m^{(\tau+1)}} \sum_{n=1}^N p(m|\mathbf{t}_n; \boldsymbol{\pi}^{(\tau)}, \mathbf{P}^{(\tau)}) \mathbf{t}_n. \tag{5}$$

The sequence of parameters obtained for $\tau = 0, 1, 2, \dots$ by iterating between the E and M steps from any starting point $\{\boldsymbol{\pi}^{(0)}, \mathbf{P}^{(0)}\}$ produces a monotonically increasing sequence of values for the log-likelihood.

A common problem of estimation in mixture distributions is that of singularities, that is, points in parameter space whose log-likelihood tends to positive infinity (e.g. a mixture of Gaussians in which one of the components is located on a data point and its variance tends to zero, thereby becoming a Dirac delta). Such singularities are undesirable because they give rise to degenerate distributions. Fortunately, the log-likelihood surface of a finite mixture of multivariate Bernoulli distributions has no singularities of value $+\infty$ (although it does have singularities of value $-\infty$). The reason is that both the log-likelihood (1) and its gradient (2)–(3) are bounded above in the whole parameter space, including its boundaries[2]. This means that estimation by the above EM algorithm from any nonpathological starting point, which is always possible by choosing p_{md} in $(0,1)$, will always lead to a proper stationary point of the log-likelihood.

3 Stationary points of the log-likelihood

We note that the second derivatives of the log-likelihood with respect to the mixing proportions are always negative (for clarity, we omit the dependence on the parameters):

$$\frac{\partial^2 l}{\partial \pi_m \partial \pi_{m'}} = -\frac{1}{\pi_m \pi_{m'}} \sum_{n=1}^{N} p(m|\mathbf{t}_n) p(m'|\mathbf{t}_n) \leq 0.$$

Therefore the Hessian of the log-likelihood has negative numbers in its diagonal and it cannot be positive definite. Hence, no stationary point of the log-likelihood is a minimum. Note that this is a general property of finite mixtures.

At any stationary point of the log-likelihood, equations (5) hold, so that we have

$$\mathrm{E}_{p(\mathbf{t})}\{\mathbf{t}\} = \sum_{m=1}^{M} \pi_m \, \mathrm{E}_{p(\mathbf{t}|m)}\{\mathbf{t}\} = \sum_{m=1}^{M} \pi_m \mathbf{p}_m = \sum_{m=1}^{M} \frac{1}{N} \sum_{n=1}^{N} p(m|\mathbf{t}_n)\mathbf{t}_n = \frac{1}{N} \sum_{n=1}^{N} \mathbf{t}_n = \bar{\mathbf{t}}$$

and so the mean of the mixture coincides with the sample mean at any stationary point. The converse does not hold generally.

The point in parameter space where $\mathbf{p}_m = \bar{\mathbf{t}}$ for all $m = 1, \dots, M$ and any distribution of the mixing proportions π_m is a stationary point of the log-likelihood, because $p(\mathbf{t}_n|m) = \prod_{d=1}^{D} \bar{t}_d^{t_{nd}} (1 - \bar{t}_d)^{(1-t_{nd})}$ is independent of m and therefore $p(m|\mathbf{t}_n) = \pi_m$ and the gradient is zero there. This point is equivalent to a single multivariate Bernoulli distribution, and should be avoided because experience shows that its log-likelihood

$$l(\{\pi_m, \mathbf{p}_m = \bar{\mathbf{t}}\}_{m=1}^{M}) = N \log \prod_{d=1}^{D} \bar{t}_d^{\bar{t}_d} (1 - \bar{t}_d)^{1-\bar{t}_d}$$

is much smaller than that of other local maxima (an intuitive fact since the mixture is trivial). Observe that a starting point of the EM algorithm in which $\mathbf{p}_m$ is the same for all components (e.g. the apparently innocuous starting point $p_{md} = 1/2$ for all m and d) will lead to the mentioned trivial mixture after one EM iteration for any original distribution of the π_m. Our experiments showed that random starting points in $(0,1)$ were less prone to leading to trivial mixtures.

4 Experimental results

4.1 Synthetic data

We generated $N = 10\,000$ vectors in a binary space of $D = 16$ dimensions from a fixed mixture of $M = 8$ 16-variate Bernoulli distributions, whose parameters are shown in figure 1. We call these the *original* parameters and denote them with an "o" superindex, e.g. $\mathbf{p}_m^{\mathrm{o}}$. From the sample alone, 10 maximum likelihood estimates were found for mixtures of $M = 4$, $M = 8$ and $M = 10$ components. We used the above EM algorithm with random starting values of the parameters p_{md} in the range $[\frac{1}{4}, \frac{3}{4}]$, stopping it when the relative change in log-likelihood was smaller than 10^{-6}. The starting values for the π_m parameters were fixed to $1/M$, thus giving each component the same weight at the beginning. The results were as follows[3]:

[2]When $p_{md} \to 0$ for some m, d, the log-likelihood gradient in equation (3) remains bounded above, because for each n, either $t_{nd} - p_{md} \to k_1 p_{md}$ (if $t_{nd} = 0$) or $p(m|\mathbf{t}_n) \propto p(\mathbf{t}_n|m) \to k_2 p_{md}$ (if $t_{nd} = 1$), where k_1 and k_2 are constants. Similarly happens with the case $p_{md} \to 1$. Therefore the log-likelihood is differentiable for $p_{md} \in [0,1]$, except in pathological situations where $|p_{md} - t_{nd}| = 1$ for all m and fixed n: in these cases $p(\mathbf{t}_n|m) = 0$ for all m and $l(\{\pi_m, \mathbf{p}_m\}_{m=1}^{M}) \to -\infty$. Since the EM algorithm always climbs the log-likelihood surface, it will not be attracted by such singularities.

[3]We quantify the distance between two vectors $\mathbf{p}$, $\mathbf{q}$ in the D-dimensional rectangle $[0,1]^D$ with the normalised undirected distance $\frac{1}{D}\left\|\mathbf{p} - \mathbf{q}\right\|_2^2$, where $\left\|\cdot\right\|_2$ is the Euclidean norm. This distance is a real number in $[0,1]$ which averages to $1/6$ for two uniformly random vectors. Observe that if $p_d = q_d + \epsilon$ for $d = 1, \dots, D$ then $\frac{1}{D}\left\|\mathbf{p} - \mathbf{q}\right\|_2^2 = |\epsilon|$ independently of D.

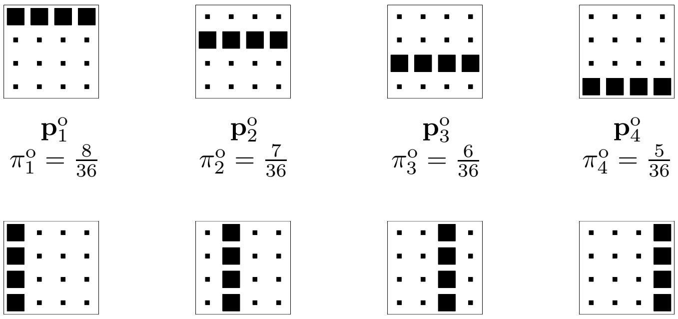

Figure 1: Parameters of the mixture of $M = 8$ 16-variate Bernoulli distributions used to generate the sample in section 4.1. For $m = 1, \dots, 8$, each 16-dimensional $\mathbf{p}_m$ vector is represented as a 4×4 image in which the area of each pixel is proportional to the value of its associated p_{md} parameter; for example, for the leftmost image, $\mathbf{p}_{1d}$ is 0.8 for $d \leq 4$ (row 1) and 0.2 for $d \geq 5$ (rows 2 to 4).

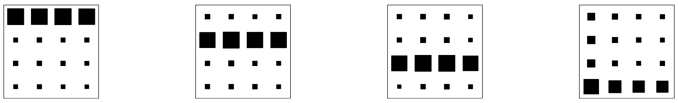

Figure 2: Parameters of a mixture of $M = 4$ 16-variate Bernoulli distributions estimated by maximum likelihood from the sample. Observe that $\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{p}_3$ and $\mathbf{p}_4$ coincide very closely with $\mathbf{p}_1^{\mathrm{o}}$, $\mathbf{p}_2^{\mathrm{o}}$, $\mathbf{p}_3^{\mathrm{o}}$ and $\frac{2}{3}\mathbf{p}_4^{\mathrm{o}} + \frac{1}{3}\mathbf{p}_5^{\mathrm{o}}$, respectively.

- Using the original number of components ($M = 8$), EM found the original parameters (both $\mathbf{p}_m$ and π_m) 9 out of 10 of the times; the normalised distance between the original and the estimated parameters was smaller than 0.0013 in those cases and the log-likelihood was $-94\,990$ (to 4 significant digits). However, the remaining estimate was a suboptimal maximum of the log-likelihood in which two of the components had the same $\mathbf{p}_m$ parameter ($\mathbf{p}_5 \approx \mathbf{p}_8 \approx \mathbf{p}_5^{\mathrm{o}}$) and a log-likelihood of $-95\,201$. The difference between the log-likelihood values of both kinds of estimates was of 0.2%.

- Using fewer components than necessary ($M = 4$) produced an estimate in which each prototype $\mathbf{p}_m$ was approximately either one of the original prototypes or a linear combination of several of the original prototypes (normalised distance smaller than 0.0023). Figure 2 shows the situation, for a particular estimate in which $\mathbf{p}_m \approx \mathbf{p}_m^{\mathrm{o}}$ for $m = 1, 2, 3$ and $\mathbf{p}_4 \approx \frac{2}{3}\mathbf{p}_4^{\mathrm{o}} + \frac{1}{3}\mathbf{p}_5^{\mathrm{o}}$. As in the previous case, estimates having different prototypes also had very close log-likelihood values (differing in 0.5%).

- Using more components than necessary ($M = 10$) always produced the 8 original $\mathbf{p}_m^{\mathrm{o}}$ vectors (normalised distance smaller than 0.0022) plus 2 extra ones, typically either repeated instances of some of the original ones or linear combinations of them. The log-likelihood value of each estimate did not differ from any of the others in more than 0.02%, indicating that once the 8 original prototypes are found, the remaining ones are largely irrelevant and reflect peculiarities of the sample used.

Experiments performed with other synthetic data sets produced the same results. We propose the following interpretation of the experimental facts. Given a large sample generated from a known mixture of multivariate Bernoulli distributions, let us construct another mixture in this way: first, pick up freely the number of components M; then, choose its $\mathbf{p}_1, \dots, \mathbf{p}_M$ prototypes either as some of the original ones or as linear combinations of them. Then, we claim that, for certain values of the mixing proportions, such a point in parameter space is very close to a maximum of the log-likelihood surface. However, we do not have theoretical support for this and we do not have a valid interpretation for the values of the mixing proportions.

The above results also suggest a procedure to follow when estimating an unknown mixture of multivariate Bernoulli distributions from a sample: choose freely a number of components M and, using EM from random

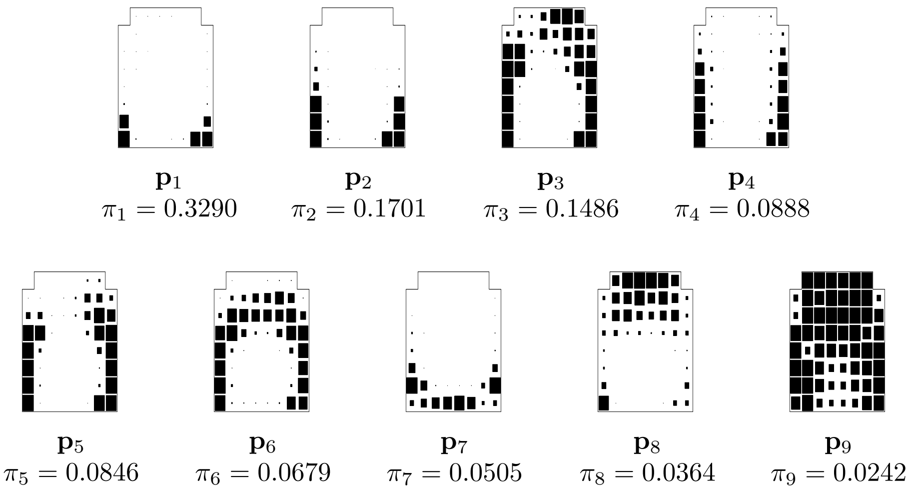

Figure 3: The 9 prototypes $\mathbf{p}_1, \dots, \mathbf{p}_9$ for a mixture of $M = 9$ multivariate Bernoulli distributions trained with the EPG data set. Each $\mathbf{p}_m$ vector, consisting of $D = 62$ values in the range $[0, 1]$, is customarily displayed as an image resembling the palate: the top row (alveolar part) contains parameters p_{m1} to p_{m6} from left to right, row 2 contains p_{m7} to p_{m14}, and so on till the bottom row (velar part), always from left to right. The area of each pixel is proportional to the value of its associated p_{md} parameter.

starting points, find several (say 10) maximum likelihood estimates for it. Inspect the prototypes obtained. If they look the same for every estimate, then M is probably the right number of components and the estimate is very close to the true generating distribution. If a fixed group of prototypes appears in each estimate, and the rest of the prototypes are repetitions of those in the group, then M is probably too big; reduce it and start again. However, if there are different prototypes in different estimates, then M is probably too small; increase it and start again.

4.2 EPG data

The technique of electropalatography (Hardcastle et al., 1989) records the presence or absence of contact between the tongue and the hard palate in a number of fixed locations of the latter and at fixed intervals during continuous speech. The result is a stream of two-dimensional binary patterns, or electropalatograms (EPGs), which can be used in speech therapy and assessment. We used a subset of electropalatography data from the EUR-ACCOR database (Marchal and Hardcastle, 1993) containing 11 852 different 62-dimensional vectors (EPGs), obtained from 12 different utterances by a native English speaker. We estimated the density of its distribution in 62-dimensional space using a finite mixture of multivariate Bernoulli distributions with $M = 6$ components. A number of estimates were found with the mentioned EM algorithm. As with the synthetic data set, the starting parameter values were $1/M$ for the π_m parameters and a random number in $\left[\frac{1}{4}, \frac{3}{4}\right]$ for the parameters p_{md}, and EM was stopped when the relative change in log-likelihood was smaller than 10^{-6}. Examination of the parameter values at this point showed that:

- Several different kinds of estimates were found, each kind being characterised by a subset of the prototypes shown in fig. 3 and by specific values for the mixing proportions. Prototypes $\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{p}_5$ and $\mathbf{p}_9$ appeared in almost every estimate (the normalised distance between corresponding prototypes did not exceed 0.02), while $\mathbf{p}_3$ and $\mathbf{p}_8$ were very common too.
- The log-likelihood of these mixtures was quite close, varying from $-138\,341$ for the combination $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_5, \mathbf{p}_7, \mathbf{p}_8, \mathbf{p}_9\}$ to $-141\,697$ for the combination $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_5, \mathbf{p}_6, \mathbf{p}_9\}$.
- Occasionally a prototype was found that could be expressed as a linear combination of some of the prototypes of fig. 3, e.g. $\frac{1}{2}\mathbf{p}_8 + \frac{1}{2}\mathbf{p}_9$.

We also found estimates for mixtures of M components, where M varied from 1 to 15. This made apparent the fact that for $M \gtrsim 9$, some prototypes appeared several times (with slight differences) in the same mixture, which therefore becomes trivial. This suggested selecting $M = 9$ as the optimum number of components for

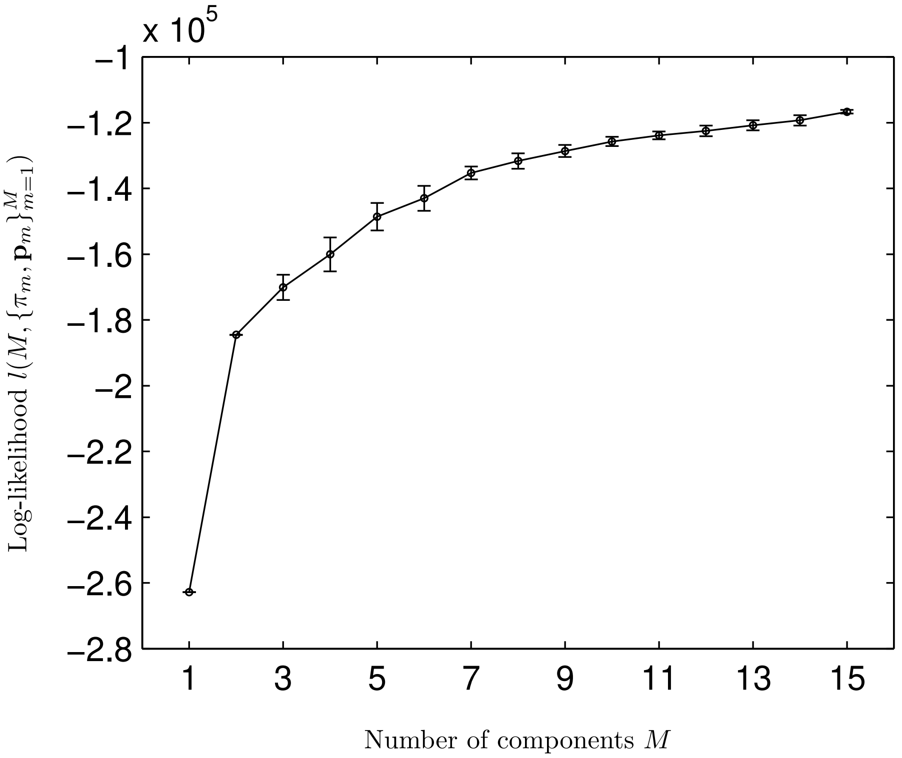

Figure 4: Log-likelihood of the mixture of multivariate Bernoulli distributions model for the EPG data as a function of the number of components M. Each value plotted corresponds to an average over 10 independent estimates.

this data set[4], with typical values for the optimal parameters (both $\mathbf{p}_m$ and π_m) given in fig. 3. Considering a sequence of mixture estimates starting from $M = 1$ to $M = 15$, prototypes with high mixture proportions tended to appear early in the sequence, although at times somewhat distorted due to the interference with other prototypes. For example, prototype $\mathbf{p}_1$ in figure 3 was present in all mixtures, while $\mathbf{p}_6$ only starts to appear (very unfrequently) for $M \geq 6$. However, prototype $\mathbf{p}_9$ appears very frequently for $M \geq 4$. Also, the log-likelihood for the data set considered increases with M and reaches a plateau for $M \approx 9$ (see fig. 4).

An interesting fact is that these prototypes are highly interpretable, corresponding to physically feasible EPGs and in fact assimilable to well-known quasi-static patterns in EPG studies (e.g. velar, alveolar, etc. (Hardcastle et al., 1989)), thus revealing important structure patterns in the data. These prototypes are similar to those produced by other methods, in particular latent variable models (Carreira-Perpiñán and Renals, 1998).

5 Conclusions

We have shown that, for the class of finite mixtures of multivariate Bernoulli distributions, the EM algorithm always converges to a proper stationary point of the log-likelihood, provided it is not started from a pathological point in parameter space (which is always possible). The reasons for this are the absence of singularities of value $+\infty$ in the log-likelihood surface and the fact that the EM algorithm always climbs that surface.

We have given empirical evidence that, for this particular class of finite mixture models, sensible and interpretable maximum likelihood estimates can be found even though that class of mixtures is not identifiable. This would suggest that the lack of identifiability might not be important from the practical point of view in some cases. Let us analyse the possible reasons with some detail.

In the first place, we note that the nonidentifiability result is not surprising if one considers that a distribution defined over a discrete domain can be specified by a finite number of equations, one for each point of

[4] An alternative way to select the critical M is to examine the log-likelihood curves for a validation set.

the domain. In our case, this means $2^D - 1$ equations (plus another one linearly dependent with them due to all the equations adding to one). Since the mixture has $MD + M - 1$ free parameters (one of the mixing proportions being linearly dependent on the others), making M larger than $\frac{2^D}{D+1}$ would yield an underdetermined system of equations with multiple solutions. However, even for small dimensions D this number is extremely large, and the number of components employed in practice will be much smaller. We restrict the rest of this discussion to a situation where the maximum number of components is much smaller than $\frac{2^D}{D+1}$, which yields an overdetermined system (where multiple solutions can still exist).

One reason for the apparent sparseness of the nonidentifiability effect may be a low population of equivalent parameter tuples. Let us call *equivalent* to two different parameter tuples $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}'$ which produce the same distribution. While we know that, for any dimension D, there exist equivalent parameter tuples, this does not mean that for each parameter tuple $\boldsymbol{\Theta}$ there will exist an equivalent, different one $\boldsymbol{\Theta}'$ (always for the case $M \ll \frac{2^D}{D+1}$). For example, it is easily seen that for the mixture in D dimensions with parameters given by:

$$\boldsymbol{\Theta} : \quad \left\{M = 2, \quad \left\{\pi_1 \in (0,1),\ \mathbf{p}_1 = (1\ 1 \dots 1)^T\right\}, \quad \left\{1 - \pi_1,\ \mathbf{p}_2 = (0\ 0 \dots 0)^T\right\}\right\},$$

which produces a distribution

$$p(\mathbf{t}|\boldsymbol{\Theta}) = \begin{cases} \pi_1, & \mathbf{t} = (1\ 1 \dots 1) \\ 1 - \pi_1, & \mathbf{t} = (0\ 0 \dots 0) \\ 0, & \text{otherwise} \end{cases},$$

there does not exist any equivalent parameter tuple $\boldsymbol{\Theta}'$ for $M < \frac{2^D}{D+1}$ (disregarding, as usual, permutations of mixture components and coincident component distributions). Thus, the actual practical problem of identifiability is: which parameter tuples $\boldsymbol{\Theta}$ have equivalent parameter tuples? Or, considering the partition of the space of parameter tuples into classes of equivalence (each class of equivalence consisting of all equivalent parameter tuples): what is the cardinality of each class? This is a difficult question to answer analytically with generality, but the experimental results suggest that nontrivial equivalence classes (consisting of more than one element) may be rare, perhaps pathological.

Another reason for the possibility of estimating a sensible parameter tuple seems to be that every estimate contains some of the original prototypes, and that the original number of components may be selected by inspection of a collection of maximum likelihood estimates obtained with the EM algorithm (as we did in section 4.1). Note that the likelihood function associated to a small sample (compared to the total number of possible different vectors, usually 2^D) need not be maximised by the original mixture.

Finally, let us remark that a reciprocal situation to the one described in this paper may also be possible, as the following example in $D = 1$ dimension shows. Consider the class of mixtures of normal distributions, which is known to be identifiable for all dimensions (Everitt and Hand, 1981). It can be readily verified that the two-component equiprobable mixture with component means at ± 1 and standard deviations of 1.5 is virtually equal to a normal of mean 0 and standard deviation 1.85 (e.g. the distance between both mixture distributions in the L_2-norm sense is 50 times smaller than the L_2-norm of any of the mixtures). A sample of enormous size would be necessary to tell one mixture from the other in terms of likelihood of the parameters. Thus, theoretical identifiability does not guarantee practical identifiability and interpretation problems may still arise.

Acknowledgments

This work was supported by ESPRIT Long Term Research Project SPRACH (20077) and by a scholarship from the Spanish Ministry of Education and Science. The first author acknowledges helpful comments from Zoubin Ghahramani.

References

Carreira-Perpiñán, M. Á. and Renals, S. (1998). Dimensionality reduction of electropalatographic data using latent variable models. *Speech Communication*, 26(4):259–282.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York.

Gyllenberg, M., Koski, T., Reilink, E., and Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *J. Appl. Prob.*, 31:542–548.

Hardcastle, W. J., Jones, W., Knight, C., Trudgeon, A., and Calder, G. (1989). New developments in electropalatography: A state-of-the-art report. *J. Clinical Linguistics and Phonetics*, 3:1–38.

Marchal, A. and Hardcastle, W. J. (1993). ACCOR: Instrumentation and database for the cross-language study of coarticulation. *Language and Speech*, 36(2, 3):137–153.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–350.