PENALIZED MAXIMUM LIKELIHOOD ESTIMATION FOR UNIVARIATE NORMAL MIXTURE DISTRIBUTIONS

A. RIDOLFI AND J. IDIER
Laboratoire des Signaux et Systèmes,
3 rue Joliot-Curie - Plateau de Moulon,
91192 Gif sur Yvette Cedex, France[†]

Abstract. Due to singularities of the likelihood function, the maximum likelihood approach for the estimation of the parameters of normal mixture models is an acknowledged ill posed optimization problem. Ill posedness is solved by penalizing the likelihood function. In the Bayesian framework, it amounts to incorporating an inverted gamma prior in the likelihood function. A penalized version of the EM algorithm is derived, which is still explicit and which intrinsically assures that the estimates are not singular. Numerical evidence of the latter property is put forward with a test.

Key words: Normal Mixtures, Maximum Likelihood, Penalized Estimator

1 Introduction

Mixture models are a well fitted tool for clustering the observations together into groups for discrimination or classification : the mixture proportions then represent the relative frequency of occurrence of each group in the population. Mixture models also provide a convenient and flexible class of models for estimating or approximating distributions.

In particular, independent identically distributed (i.i.d.) mixture models well fit several problems in signal and image processing, covering a wide range of applications. In [1] a Bernoulli-Gaussian mixture model is adopted in a deconvolution problem, while [2] highlights the important role of mixture models in the field of cluster analysis. An example of the application of mixtures in biological (plant morphology measures) and physiological (EEG signals) data modeling is presented in [3]. Markovian mixture models are also commonly used, as in [4] where an application to medical image segmentation is considered.

The present contribution summarizes two of our previous works [5, 6], which focus on i.i.d. mixtures of univariate normal densities. Parameters are estimated

[†]Email: ridolfi@lss.supelec.fr, idier@lss.supelec.fr

with a penalized maximum likelihood approach, by mean of the EM algorithm [7].

2 Mixture model

We consider a sample $\boldsymbol{x} = \{x_1, \ldots, x_T\}$ of an i.i.d. mixture of N univariate normal densities

$$f\left(x;\boldsymbol{\theta}\right) = \sum_{i=1}^{N} a_i f\left(x;\mu_i,\sigma_i^2\right)$$

where $f\left(x;\mu_i,\sigma_i^2\right) = \left(2\pi\sigma_i^2\right)^{-\frac{1}{2}} \exp\left\{-\left(x_k-\mu_i\right)^2/2\sigma_i^2\right\}$. $\boldsymbol{\theta} = \{\theta_1,\ldots,\theta_N\}$ are the mixture parameters, belonging to the parameter space

$$\Theta = \left\{ \theta_i = \left\{a_i,\mu_i,\sigma_i^2\right\} \mid a_i \in \mathbb{R}_+,\ \sum_{i=1}^{N} a_i = 1\ ; \right.$$
$$\left. \sigma_i^2 \in \mathbb{R}_+/\left\{0\right\}\ ;\ \mu_i \in \mathbb{R} \text{ for } i = 1,\ldots,N \right\}$$

Given $\boldsymbol{x}$, the maximum likelihood estimate of the mixture parameters is defined as:

$$\widehat{\boldsymbol{\theta}}_T \mid f(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_T) = \sup_{\boldsymbol{\theta}\in\Theta} f\left(\boldsymbol{x};\boldsymbol{\theta}\right) \tag{1}$$

where

$$f\left(\boldsymbol{x};\boldsymbol{\theta}\right) = \prod_{k=1}^{T} f\left(x_k;\boldsymbol{\theta}\right) = \prod_{k=1}^{T}\sum_{i=1}^{N} a_i f\left(x_k;\mu_i,\sigma_i^2\right) \tag{2}$$

is the likelihood function.

3 Likelihood function degeneracy

Likelihood function degeneracy toward infinity is a well known problem for mixtures of Gaussian distributions, first put forward with a simple example in [8] (see also [9]). Such an example considered a two class mixture model with a corresponding likelihood function given by

$$f\left(\boldsymbol{x};\boldsymbol{\theta}\right) = \prod_{k=1}^{T}\left(a_1\left(\sigma_1^2\right)^{-\frac{1}{2}} \exp\left\{-\left(x_k-\mu_1\right)^2/2\sigma_1^2\right\} \right.$$
$$\left. + a_2\left(\sigma_1^2\right)^{-\frac{1}{2}} \exp\left\{-\left(x_k-\mu_1\right)^2/2\sigma_1^2\right\}\right)\left(2\pi\right)^{-\frac{1}{2}} \tag{3}$$

Intuitively, the degeneracy is due to the fact that in the sum of Gaussian densities the variance parameter appears in the denominator. Indeed, couples such as $(\sigma_i^2 = 0, \mu_i = x_k)$ yield singularities, in the sense that f tends to infinity as $\boldsymbol{\theta}$ approaches one of the corresponding points, located at the boundary of Θ, as rigorously stated by the following property.

Property 3.1 *Let us consider the likelihood function (2), then*

$$\forall \boldsymbol{x} \in \mathbb{R}^T,\ \exists\ \boldsymbol{\theta}^0 \in \bar{\Theta} \mid \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^0} f\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right) = +\infty$$

where Θ *is the parameter space,* $\bar{\Theta}$ *is the closure of the parameter space,* $\boldsymbol{\theta}^0 = \left\{\boldsymbol{a}, \boldsymbol{\mu} = x_k, {\boldsymbol{\sigma}^2}^0 = 0\right\} \in \bar{\Theta}$ *is a point in the closure of the parameter space, and* $\boldsymbol{\theta} = \left\{\boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2\right\} \in \Theta$ *is a point in the parameter space.*

Consequently, the maximum likelihood estimator (1) cannot be defined. In practice, unboundedness of $f\left(\boldsymbol{x}; \boldsymbol{\theta}\right)$ is a cause of failure of commonly used optimization algorithms, for instance of EM [9] and gradient types.

We will specifically refer to the EM algorithm, which iteratively compute the maximum likelihood estimates by mean of the following re-estimation formulas

$$a_i^{j+1} = M\left(\boldsymbol{\theta}^j\right)/T \tag{4}$$

$$\mu_i^{t+1} = \frac{\sum_{k=1}^T x_k \frac{a_i^j f\left(x_k; \mu_i^j, {\sigma_i^2}^j\right)}{f\left(x_k; \boldsymbol{\theta}^j\right)}}{M\left(\boldsymbol{\theta}^j\right)} \tag{5}$$

$$\sigma_i^{2\,j+1} = \frac{\sum_{k=1}^T \left(x_k - \mu_i^j\right)^2 \frac{a_i^j f\left(x_k; \mu_i^j, {\sigma_i^2}^j\right)}{f\left(x_k; \boldsymbol{\theta}^j\right)}}{M\left(\boldsymbol{\theta}^j\right)} \tag{6}$$

where $M\left(\boldsymbol{\theta}^j\right) = \sum_{k=1}^T a_i^j f\left(x_k; \mu_i^j, {\sigma_i^2}^j\right)/f\left(x_k; \boldsymbol{\theta}^j\right)$ and j indicates the iteration.

4 Bayesian solution to degeneracy: penalized likelihood function

A Bayesian solution is proposed to solve the degeneracy of the likelihood function in the origin of any of the variance parameters. The latter are considered as i.i.d. random variables, leading to a penalized likelihood function

$$f_{\mathrm{P}}\left(\boldsymbol{x}, \boldsymbol{\theta}\right) = f\left(\boldsymbol{x}, \boldsymbol{\sigma}^2; \boldsymbol{a}, \boldsymbol{\mu}\right) = f\left(\boldsymbol{x} | \boldsymbol{\sigma}^2; \boldsymbol{a}, \boldsymbol{\mu}\right) \prod_{i=1}^N g\left(\sigma_i^2\right) \tag{7}$$

where g is the common prior probability density of variance parameters.

Our goal is to adjust g so that the penalized likelihood is a bounded function that can be locally maximized by mean of an EM algorithm (which can be referred to as a "penalized" EM algorithm). In other words, g must satisfy the requirements of

1. being a proper probability density function,
2. tending appropriately to zero to compensate for the likelihood singularities,
3. and allowing to maintain explicit re-estimation formulas for the resulting penalized EM algorithm.

The inverted gamma distribution

$$g\left(\sigma_i^2\right) = \frac{\alpha^{\beta-1}}{\Gamma\left(\beta-1\right)} \frac{1}{\sigma_i^{2\beta}} \exp\left\{-\frac{\alpha}{\sigma_i^2}\right\} 1_{[0,+\infty)} \tag{8}$$

where $i = 1 \dots N$, is proved to satisfy the three conditions.

On the other hand, the inverted gamma distribution is known as the *conjugate prior* for the variance of a scalar Gaussian density [10].

As regard Point 1, the inverted gamma is assured to be proper by constraining the choice of its parameters: $\alpha > 0$ and $\beta > 1$, as discussed in [10].

As regard Point 2, the following property states the boundedness of f_{P} on Θ (whereas, from Property 3.1, f is an unbounded function under the same conditions), and it assures that the points of singularity do not maximize f_{P}.

Property 4.1 *The penalized likelihood is bounded above over the parameters space. Hence, the penalized likelihood function does not degenerate in any point of the closure of parameters space $\bar{\Theta}$. Moreover it tends to zero as $\sigma^2 \to 0$. Hence, no $\sigma_i^2 = 0, i \in \{1 \dots N\}$ maximizes the penalized likelihood function.*

Proof 4.1 For the sake of simplicity the proof refers to a two class mixture model, without loss of generality.

Akin to the likelihood function, the penalized version (7) may degenerate only in the origin of any of the parameters $\boldsymbol{\sigma}^2$. Let us note $K = (2\pi)^{-\frac{T}{2}} \alpha^{2(\beta-1)} / \Gamma\left(\beta-1\right)^2$, and let us consider the likelihood function (3) penalized by a proper inverted gamma distribution (7)

$$f_{\mathrm{P}}\left(\boldsymbol{x}, \boldsymbol{\theta}\right) = K \frac{1}{\sigma_1^{2\beta}} \exp\left\{-\frac{\alpha}{\sigma_1^2}\right\} \frac{1}{\sigma_2^{2\beta}} \exp\left\{-\frac{\alpha}{\sigma_2^2}\right\} \prod_{k=1}^{T} \left(\frac{a_1}{{\sigma_1^2}^{\frac{1}{2}}} \exp\left\{-\frac{\left(x_k-\mu_1\right)^2}{2\sigma_1^2}\right\} + \frac{a_2}{{\sigma_2^2}^{\frac{1}{2}}} \exp\left\{-\frac{\left(x_k-\mu_2\right)^2}{2\sigma_2^2}\right\} \right)$$

On every compact domain contained in the parameter space, f_{P} is bounded. This is a straightforward consequence of the fact that f_{P} is the product of two functions which are bounded on such domains (the product of sum of gaussian distributions and the product of inverted gamma distributions). Hence, it is sufficient to prove that f_{P} remains bounded on the boundaries of Θ, and more precisely that it remains bounded in the points of singularity.

From the inequality $\exp\left\{-\frac{\left(x_k-\mu_1\right)^2}{2\sigma_1^2}\right\} \leq 1$ the likelihood function can be bounded

above by

$$\leq K\frac{1}{{\sigma_1}^{2\beta}}\exp\left\{-\frac{\alpha}{\sigma_1^2}\right\}\frac{1}{{\sigma_2}^{2\beta}}\exp\left\{-\frac{\alpha}{\sigma_2^2}\right\}\prod_{k=1}^{T}\left(\frac{a_1}{{\sigma_1^2}^{\frac{1}{2}}}+\frac{a_2}{{\sigma_2^2}^{\frac{1}{2}}}\right)$$

$$= K\prod_{k=1}^{T}\left(\frac{1}{{\sigma_1^2}^{\frac{\beta}{T}+\frac{1}{2}}}\exp\left\{-\frac{\alpha}{T\sigma_1^2}\right\}\frac{1}{{\sigma_2^2}^{\frac{\beta}{T}}}\exp\left\{-\frac{\alpha}{T\sigma_2^2}\right\}\right.$$

$$\left.+\frac{1}{{\sigma_1^2}^{\frac{\beta}{T}}}\exp\left\{-\frac{\alpha}{T\sigma_1^2}\right\}\frac{1}{{\sigma_2^2}^{\frac{\beta}{T}+\frac{1}{2}}}\exp\left\{-\frac{\alpha}{T\sigma_2^2}\right\}\right) \quad (9)$$

By considering that

$$\lim_{\sigma^2\to 0}\frac{1}{{\sigma^2}^{\frac{\beta}{T}+\frac{1}{2}}}\exp\left\{-\frac{\alpha}{T\sigma^2}\right\}=0$$

and that

$$\lim_{\sigma^2\to 0}\frac{1}{{\sigma^2}^{\frac{\beta}{T}}}\exp\left\{-\frac{\alpha}{T\sigma^2}\right\}=0$$

it is straightforward to see that the penalized likelihood function tends to zero as $\sigma^2 \to 0$. Therefore, it is bounded in the point of singularity and its boundedness on the whole parameter space follows. □

Therefore, the existence of the penalized maximum likelihood estimator is assured, and such an estimator falls within the parameters space Θ (the boundaries are excluded by the null value of the likelihood).

Moreover, the penalized likelihood estimator has recently been proved to be consistent [6].

5 Penalized EM algorithm

As regard Point 3, explicitness directly follows from constrained adjustment of g. However, a more thorough analysis reveals that the re-estimation equations remain explicit *because* g is chosen as the *conjugate prior* of the likelihood of the *complete data*.

Indeed, the EM algorithm is based on the maximization of a criterion Q which depends indirectly on the likelihood function and which guarantees the maximization of the latter. Explicitness of the re-estimation equations is related to the form of the terms contained in such a criterion. In the case of mixture models, one of these terms is the likelihood function of the complete data (*i.e.,* $f\left(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{a},\boldsymbol{\mu},\boldsymbol{\sigma}^2\right)$ where $\boldsymbol{c}$ indicates to which class belongs each element x_i of the sample $\boldsymbol{x}$). By applying a penalization, such a term changes to $f\left(\boldsymbol{x},\boldsymbol{\sigma}^2|\boldsymbol{c},\boldsymbol{a},\boldsymbol{\mu}\right)$, becoming proportional to the *a posteriori* likelihood of the complete data. On the other hand, the conjugate prior $g(\theta)$ of a distribution $f(x|\theta)$ is, by definition (see [10]), the prior that gives an *a posteriori* distribution $f(\theta|x)$ belonging to its same family. Moreover, in the case of gaussian mixtures, $f\left(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{a},\boldsymbol{\mu},\boldsymbol{\sigma}^2\right)$ and $g\left(\boldsymbol{\sigma}^2\right)$ have,

with respect to $\boldsymbol{\sigma}^2$, the same structure. Hence, by substituting $f\left(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{a},\boldsymbol{\mu},\boldsymbol{\sigma}^2\right)$ with $f\left(\boldsymbol{x},\boldsymbol{\sigma}^2|\boldsymbol{c},\boldsymbol{a},\boldsymbol{\mu}\right)$, no "structural" changes are made and the explicitness is maintained.

The re-estimation equations of the penalized EM algorithm are not only explicit, but they also correspond to a very slight alteration of the standard ones. Indeed, equations (4) and (5) remain unchanged, while equation (6) becomes

$$\sigma_i^{2\,j+1} = \frac{2\alpha + \sum_{k=1}^{T}\left(x_k - \mu_i^j\right)^2 \frac{a_i^j f\left(x_k;\mu_i^j,\sigma_i^{2\,j}\right)}{f\left(x_k;\boldsymbol{\theta}^j\right)}}{2\beta + M\left(\boldsymbol{\theta}^j\right)} \tag{10}$$

Therefore, penalization of the EM does not increase the computational burden: this is an extremely important aspects in the case of large samples or image processing.

Moreover, from equation (10) it is straightforward to see that every maximizer (either global or local) of the penalized likelihood function yields strictly positive variance estimates $\hat{\sigma}_i^2 \geq \sigma_{\min}^2(T) > 0$, where $\sigma_{\min}^2(T)$ tends to 0 as T tends to infinity.

6 Numerical results

We have tested the penalized and non penalized EM algorithm on a 2 class mixture model, defined in (3).

Eight-hundred samples of length fifty have been randomly generated from two gaussian distribution, having parameters $\boldsymbol{a} = [0.5\ 0.5]\,, \boldsymbol{\mu} = [0\ 2.5]\,, \boldsymbol{\sigma}^2 = [1\ 2]$.

For each sample, the starting point of the EM iterations was chosen automatically. Such a choice is based on partioning the empirical histogram of the data, as proposed in [11]. As in [12], the EM algorithm was considered to have converged whenever the maximum of the relative stepsize

$$|{a_i}^{j+1} - {a_i}^j|/{a_i}^j, |{\mu_i}^{j+1} - {\mu_i}^j|/{\mu_i}^j, |\sigma_i^{2\,j+1} - \sigma_i^{2\,j}|/\sigma_i^{2\,j}$$

for $i = 1 \ldots N$, became less or equal than 10^{-5}.

Figure 1 depicts the histograms for the values of the non-penalized estimates of σ_1^2 and σ_2^2, and the histograms for the values of the penalized ones. By comparing the histograms, the efficiency of penalization becomes evident. Without penalization, the distribution of the estimates spreads toward the singularity ($\sigma^2 = 0$, hence $\log\sigma^2 = -\infty$), and for 13 times the EM algorithm converges to the singularity itself. On the other hand, coherently with the theoretical results of Property 4.1, the estimates computed by the penalized EM algorithm are concentrated around the true value and none of them is a singularity.

By increasing the length of the samples the number of convergence of the standard EM algorithm to singularities is reduced (probably as a consequence of a restriction of the attracting domain of the degeneracy point), but it is still greater than zero. Table 1 summarizes the results for samples of length fifty and one-hundred of the non-penalized (a) and penalized (b) EM algorithm.

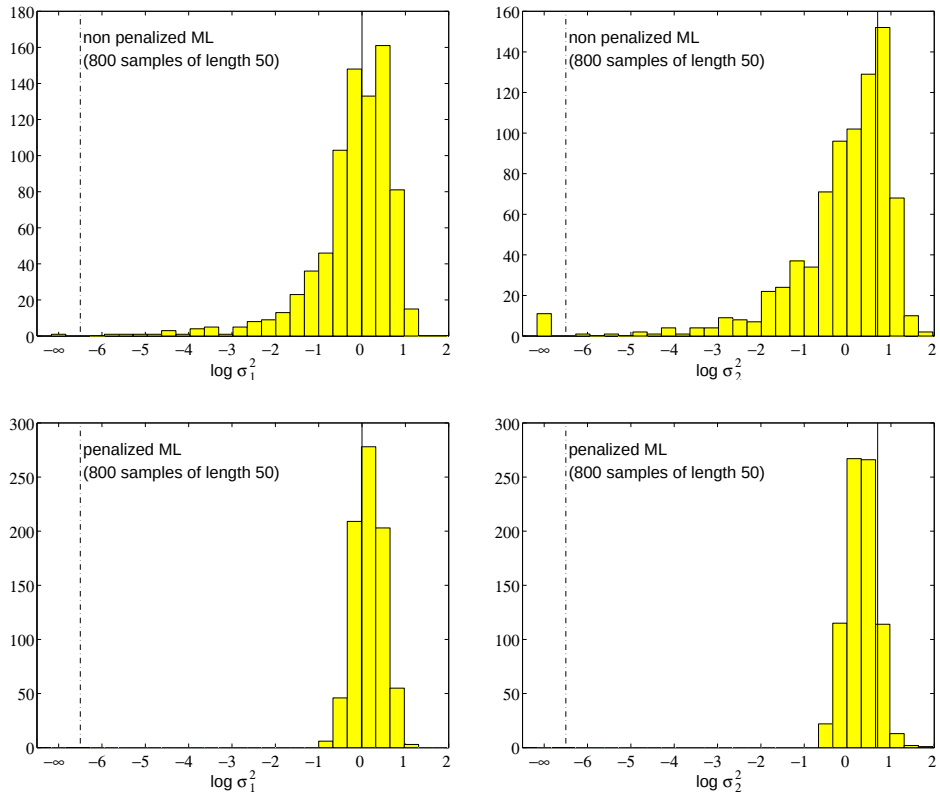

Figure 1: histograms of EM σ_1^2 and σ_2^2 estimates, where the solid line indicates the true value while the dashed line indicates a rupture toward infinity of the x axis. The top two histograms and the bottom ones refer to the values of the non-penalized and penalized estimates, respectively. Penalization evidently avoid spreading toward the singularity ($\sigma^2 = 0$, hence $\log \sigma^2 = -\infty$) of the σ^2 estimates.

Table 1: non penalized (a) and penalized (b) EM algorithm results for samples of length fifty and one-hundred.

(a)

800 samples of length:	convergence to singularities:
50	13
100	1

(b)

800 samples of length:	min value of σ^2:
50	0.3951
100	0.4247

7 Concluding remarks

Penalization of the likelihood has revealed itself to be an efficient and simple solution to likelihood degeneracy.

Theoretical properties assured the existence of the maximum likelihood estimator as well as its belonging to the parameter space.

The choice of the conjugate prior of the likelihood of the complete data as penalization term conducted to explicit EM algorithm re-estimation formulas. While the role of conjugate priors is acknowledged in Bayesian sampling schemes, including in mixture problems [13], putting forward the link between conjugate priors and explicit penalized EM schemes is an original contribution, as far as we know.

Numerical examples put in evidence the existence of the singularities and the efficiency of the penalized solution.

Concerning the asymptotic behavior of the penalized maximum likelihood estimate, we know from [14] that the penalization does not alter asymptotic properties such as consistency. Hence, local consistency of the penalized estimate is a direct consequence of local consistency of the non penalized one (see [14]). On the other hand, global consistency cannot be similarly deduced, since non penalized maximum likelihood estimate is globally not even defined and classical theorems, as [15] and [8], cannot be applied. Although not trivial, proof of global consistency has recently been achieved [6].

To our best knowledge, Hathaway's EM re-estimation formulas [12] are the only preexisting non-degenerate alternative to our penalized version. It is based on constrained maximization of the likelihood, within an appropriately chosen subset of Θ. However, Hathaway's version is substantially more complex to derive and to implement, and the resulting numerical cost is higher.

References

[1] F. Champagnat, Y. Goussard, and J. Idier, "Unsupervised deconvolution of sparse spike trains using stochastic approximation," *IEEE Trans. Signal Processing*, **44**, pp. 2988–2998, Dec. 1996.

[2] G. J. McLachlan and K. E. Basford, *Mixture Models, inference and applications to clustering*, vol. 84 of *statistics*, Dekker, 1987.

[3] S. J. Roberts, D. Husmeier, I. Rezek, and W. Penny, "Bayesian approaches to Gaussian mixture modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**, pp. 887–906, Nov. 1998.

[4] A. Ridolfi, *Maximum Likelihood Estimation of Hidden Markov Model Parameters, with Application to Medical Image Segmentation.* Tesi di Laurea, Politecnico di Milano, Facoltà di Ingegneria, Milano, Italia, 1997.

[5] A. Ridolfi and J. Idier, "Penalized maximum likelihood estimation for univariate normal mixture distributions," in *Actes du 17e colloque GRETSI*, (Vannes, France), pp. 259–262, Sept. 1999.

[6] G. Ciuperca, A. Ridolfi, and J. Idier, "Penalized maximum likelihood estimator for normal mixtures," tech. rep., 2000.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. B*, **39**, pp. 1–38, 1977.

[8] J. Kiefer and J. Wolfowitz, "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," *Ann. Math. Statist.*, **27**, pp. 887–906, 1956.

[9] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, **26**, pp. 195–239, Apr. 1984.

[10] C. Robert, *L'analyse statistique Bayésienne*, Economica, 1992.

[11] P. A. Devijver and M. Dekessel, "Champs aléatoires de Pickard et modélisation d'images digitales," *Traitement du Signal*, **5**, (5), pp. 131–150, 1988.

[12] R. J. Hathaway, "A constrained EM algorithm for univariate normal mixtures," *J. Statist. Comput. Simul.*, **23**, pp. 211–230, 1986.

[13] J. Diebolt and C. P. Robert, "Estimation of finite mixture distributions through Bayesian sampling," *J. R. Statist. Soc. B*, **56**, (2), pp. 363–375, 1994.

[14] R. A. Redner, "Maximum likelihood estimation for mixture models," technical memorandum, NASA, Oct. 1980.

[15] A. Wald, "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Stat.*, **20**, pp. 595–601, 1949.