# Robust $L_2E$ Parameter Estimation of Gaussian Mixture Models: Comparison with Expectation Maximization

Umashanger Thayasivam[1], Chinthaka Kuruwita[2], and Ravi P. Ramachandran[1(✉)]

[1] Rowan University, Glassboro, NJ, USA
{thayasivam,ravi}@rowan.edu
[2] Hamilton College, Clinton, NY, USA
ckuruwit@hamilton.edu

**Abstract.** The purpose of this paper is to discuss the use of $L_2E$ estimation that minimizes integrated square distance as a practical robust estimation tool for unsupervised clustering. Comparisons to the expectation maximization (EM) algorithm are made. The $L_2E$ approach for mixture models is particularly useful in the study of big data sets and especially those with a consistent numbers of outliers. The focus is on the comparison of $L_2E$ and EM for parameter estimation of Gaussian Mixture Models. Simulation examples show that the $L_2E$ approach is more robust than EM when there is noise in the data (particularly outliers) and for the case when the underlying probability density function of the data does not match a mixture of Gaussians.

**Keywords:** Robust $L_2E$ estimation · Gaussian mixture model · Expectation maximization · Unsupervised learning · Big data

## 1 Introduction and Motivation

Mixture models and in particular Gaussian Mixture Models (GMM), are commonly used for density estimation and classification. In this era of Big Data and everyday, the data is highly complex and enormous in size. Mixture models offer a powerful and flexible way to represent the data. A comprehensive discussion on mixture models can be found in [1,2].

When the number of mixture components is known and the component densities are assumed to belong to a specified parametric family, the popular Expectation Maximization (EM) algorithm [3] based on Maximum Likelihood Estimation (MLE) is often used to estimate the GMM parameters. However, when there is a small perturbation in one of the component densities, MLE becomes significantly biased and very sensitive to outliers [4]. Furthermore, when the data is not Gaussian, the EM method may not cluster a set of data points to a Gaussian with a meaningful mean vector and covariance matrix. The EM based approach


S. Arik et al. (Eds.): ICONIP 2015, Part III, LNCS 9491, pp. 281–288, 2015.
DOI: 10.1007/978-3-319-26555-1_32

is not robust when the underlying probability density function of the data does not match a mixture of Gaussians (known as a data/model mismatch).

To overcome this limitation, Scott [5–8] introduced an alternative minimum distance estimation method based on the integrated squared error criterion (termed $L_2E$) which avoids the use of nonparametric kernel density estimators. The $L_2E$ approach is a special case of a general method introduced in [9] that is based on a whole continuum of divergence estimators that begin with MLE and interpolate to the $L_2E$ estimator. Markatou [10] used the weighted likelihood estimation approach to address the effects of data/model mismatch on parameter estimates.

In this paper, the focus is on the $L_2E$ as an alternative to the EM for parameter estimation of models with a known finite number of mixtures. A discussion of the EM and $L_2E$ approaches are given. Simulation results specific to GMM are shown to depict the robustness property of the $L_2E$ method with respect to noise in the data (particularly outliers) and data/model mismatch [11–13].

The basic notation in this paper is as follows. Let $f_{\boldsymbol{\theta}_m}(x)$ denote a general mixture probability density function with $m$ components as given by

$$f_{\boldsymbol{\theta}_m}(\boldsymbol{x}) = \sum_{i=1}^{m} \pi_i f(\boldsymbol{x}|\boldsymbol{\phi}_i) \tag{1}$$

where $\boldsymbol{\theta}_m = (\pi_1, \ldots, \pi_{m-1}, \pi_m, {\boldsymbol{\phi}_1}^T, \ldots, {\boldsymbol{\phi}_m}^T)^T$, the weights $\pi_i > 0$, $\sum_{i=1}^{m} \pi_i = 1$ and $f(\boldsymbol{x}|\boldsymbol{\phi}_i)$ is a probability density function with parameter vector $\boldsymbol{\phi}_i$. In theory, the $f(\boldsymbol{x}|\boldsymbol{\phi}_i)$ could be any parametric density, although in practice they are often from the same parametric family (usually Gaussian).

## 2 EM Algorithm

The Expectation-Maximization (EM) algorithm [3] is broadly based on the iterative computation of MLE. The EM method alternates between two steps:

1. Expectation (E) step: Computes an expectation of the likelihood by including the latent variables as if they were observed and a
2. Maximization (M) step: computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found in the E step.

The parameters found in the M step are then used to begin another E step and the process is repeated.

For finite mixture models, the observed data samples $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ are viewed as incomplete. The complete data is obtained as $\boldsymbol{Z} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ for $i = 1$ to $n$ where $\boldsymbol{y}_i = (\boldsymbol{y}_{1i}, \cdots, \boldsymbol{y}_{mi})^T$ is a latent (unobserved or missing) indicator vector with $\boldsymbol{y}_{ij} = 1$ if $\boldsymbol{x}_i$ is from the mixture component j and zero otherwise. The log-likelihood of $\boldsymbol{Z}$ is defined by

$$L(\boldsymbol{\theta}_m|\boldsymbol{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} \log y_{ij} \log[\pi_j f(\boldsymbol{x}_i|\boldsymbol{\phi}_j)] \tag{2}$$

The EM algorithm obtains a sequence of estimates $\boldsymbol{\theta}^{(t)}, t = 0, 1, \cdots$ by alternating the E-Step and the M-Step until some convergence criterion is met.

1. **E-Step:** Calculate the Q function, the conditional expectation of the complete log-likelihood, given $\boldsymbol{X}$ and the current estimate $\boldsymbol{\theta}^{(t)}$.
2. **M-Step:** Update the estimate of the parameters by maximizing the Q function.

In the case of GMM, maximizing Q provides an explicit solution. In most instances, EM has the advantages of reliable global convergence, low cost per iteration, economy of storage, ease of programming and heuristic appeal. However, its convergence can be very slow in simple problems which are often encountered in practice. Also, when there is a small perturbation in one of the component densities due to noise in the data, the MLE estimates become highly unstable due to the lack of robustness to outliers. For the case of GMM [14], this can be seen easily as maximization of the likelihood function under an assumed Gaussian distribution is equivalent to finding the least-squares solution, whose lack of robustness is well known. As a robust alternative we discuss an approach based on the minimization of the integrated square distance, namely $L_2E$.

## 3 Robust $L_2E$ Estimator

The integrated squared distance has been used as the goodness-of-fit criterion in nonparametric density estimation for a long time. In the classic papers of Scott [6,7], an alternative minimum distance estimation method based on the integrated squared error criterion, termed $L_2E$, was introduced and has the following attributes.

1. The use of nonparametric kernel density estimators is avoided.
2. The $L_2E$ is especially suited for parameter-rich models such as mixture models.
3. The genesis of Scott the $L_2E$ approach, which can be traced to the pioneering work of Rudemo [15] and Bowman [16], is computationally feasible and leads to robust estimators.
4. The $L_2E$ is a special class of robust estimators like the median-based estimators, which sacrifice some asymptotic efficiency for substantial computational benefits in difficult estimation problems.
5. The $L_2E$ estimator performs much better than other robust estimators such as minimum Hellinger estimates ($MHD$) under severe data contamination.

The $L_2E$ estimator belongs to the family of minimum density power divergence ($MDPD$) estimators introduced in [9] with the tuning parameter $\alpha = 1$. The tuning parameter $\alpha$ in an $MDPD$ estimator controls the trade-off between robustness and efficiency. It is also shown that the robustness of the $L_2E$ estimator is achieved at a fairly stiff price in asymptotic efficiency [9]. For the normal, exponential and Poisson distributions with small values of $\alpha \leq 0.10$, the $MDPD$ has strong robustness properties and retains high asymptotic relative efficiency

(ARE) with respect to MLE. However, within the family of density-based power divergence measures, the $L_2E$ approach has the distinct advantage that a key integral can be computed in closed form, especially for Gaussian mixtures.

### 3.1 $L_2E$ Algorithm

Given the true probability density $g(\boldsymbol{x})$ and the finite mixture with $m$ components, $f_{\boldsymbol{\theta}_m}(\boldsymbol{x})$, consider the $L_2$ distance between $f_{\boldsymbol{\theta}_m}$ and $g(\boldsymbol{x})$ as given by

$$L_2(f_{\boldsymbol{\theta}_m}, g(\boldsymbol{x})) = \int_{-\infty}^{\infty} [f_{\boldsymbol{\theta}_m}(\boldsymbol{x}) - g(\boldsymbol{x})]^2 d\boldsymbol{x}. \tag{3}$$

The aim is to derive an estimate of $\boldsymbol{\theta}_m$ that minimizes the $L_2$ distance [5–7,11–13]. Expanding Eq. (3) gives

$$\begin{aligned} L_2(f_{\boldsymbol{\theta}_m}, g(\boldsymbol{x})) = &\int_{-\infty}^{\infty} f_{\boldsymbol{\theta}_m}^2(\boldsymbol{x}) d\boldsymbol{x} - 2\int_{-\infty}^{\infty} f_{\boldsymbol{\theta}_m}(\boldsymbol{x}) g(\boldsymbol{x}) d\boldsymbol{x} \\ &+ \int_{-\infty}^{\infty} g(\boldsymbol{x})^2 d\boldsymbol{x} \end{aligned} \tag{4}$$

where the last integral is a constant with respect to $\boldsymbol{\theta}_m$ and therefore, may be ignored for the minimization. The first integral in Eq. (4) is often available as a closed form expression that, for Gaussian mixtures, may be evaluated for any specified value of $\boldsymbol{\theta}_m$ as shown later in Eq. (7). The second integral in Eq. (4) is simply the average height of the density estimate, which may be estimated as $-2n^{-1}\sum_{i=1}^{n} f_{\boldsymbol{\theta}_m}(\boldsymbol{X}_i)$ where $\boldsymbol{X}_i$ is a sample observation. Based on the above analysis, the $L_2E$ estimator of $\boldsymbol{\theta}_m$ is given by

$$\hat{\boldsymbol{\theta}}_m^{L_2E} = \arg\min_{\boldsymbol{\theta}_m} \left[ \int_{-\infty}^{\infty} f_{\boldsymbol{\theta}_m}^2(\boldsymbol{x}) d\boldsymbol{x} - 2n^{-1} \sum_{i=1}^{n} f_{\boldsymbol{\theta}_m}(\boldsymbol{X}_i) \right], \tag{5}$$

### 3.2 GMM Models

For multivariate Gaussian mixtures,

$$f(\boldsymbol{x}|\boldsymbol{\phi}_i) = \phi(\boldsymbol{x}|\ \boldsymbol{\mu}_i, \Sigma_i) \tag{6}$$

where $\boldsymbol{\mu}_i$ is the mean vector and $\Sigma_i$ is the covariance matrix for component $i$. In this case, the problem reduces to finding the $L_2E$ estimator for a Gaussian Mixture Model (GMM). Now, the first integral in Eq. (4) reduces to

$$\int_{-\infty}^{\infty} f_{\boldsymbol{\theta}_m}^2(\boldsymbol{x}) d\boldsymbol{x} = \sum_{k=1}^{m}\sum_{l=1}^{m} \pi_k \pi_l \ \ \phi(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l|\ 0, \Sigma_k + \Sigma_l), \tag{7}$$

thereby making Eq. (4) tractable for minimization and significantly reducing the computations involved in getting the $L_2E$ estimator. Since this is a computationally feasible closed-form expression, estimation of the GMM parameters by the $L_2E$ procedure may be performed by any standard nonlinear optimization algorithm [5,6,11–13]. In this work, we used the *'nlminb'* nonlinear minimization routine in [17].

# 4 Experimental Results

## 4.1 Performance Due to Data Contamination (Outliers)

In this section, simulations using EM and $L_2E$ parameter estimates are compared when there is no data contamination and when there is (with and without the presence of outliers/noise).

**Gaussian Mixture Model with No Outliers:** A GMM model $f(x)$ with two components, each being a univariate Gaussian density $\phi(x)$ is simulated as given by

$$f(x) = 0.75\phi(x|\ \mu_1 = 0, \sigma_1^2 = 1) + 0.25\phi(x|\ \mu_2 = 1, \sigma_2^2 = 1). \tag{8}$$

The variable $\mu$ denotes the mean and the variable $\sigma^2$ denotes the variance. A total of 10000 sample points from the above Gaussian mixture (see Eq. (8)) are generated and parameter estimation is performed. A total of 100 Monte Carlo simulations are performed to evaluate consistency and efficiency.

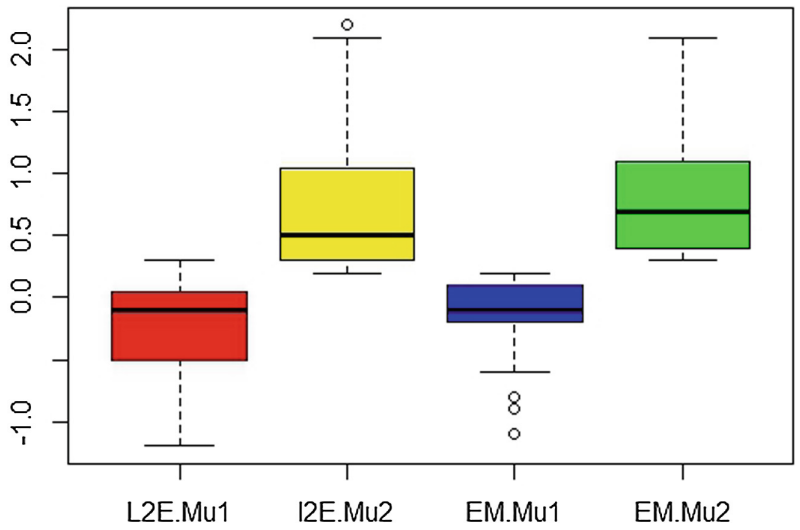


**Fig. 1.** Boxplots of the estimated mean for $L_2E$ and EM from 100 Monte Carlo Simulations of a GMM Model With No Outliers

The boxplots of the parameter estimates of the component means for the mixture model in Eq. (8) with no data contamination are shown in Fig. 1. The results clearly show that both solutions are comparable and close to the true estimates. Note that the average of the 100 Monte Carlo estimates of the $L_2E$ and EM means are close to the true value.

**Gaussian Mixture Model with Outliers:** The second simulation extends our study by adding outliers to illustrate the robustness property of $L_2E$ against outliers. In this case, 9900 sample points from the above Gaussian mixture in Eq. (8) are contaminated by adding 100 sample points (outliers) simulated from $\phi(x|\ \mu = 5, \sigma^2 = 1)$. Once again, 100 Monte Carlo simulations are performed to evaluate the performance of $L_2E$ and EM for consistency and efficiency.

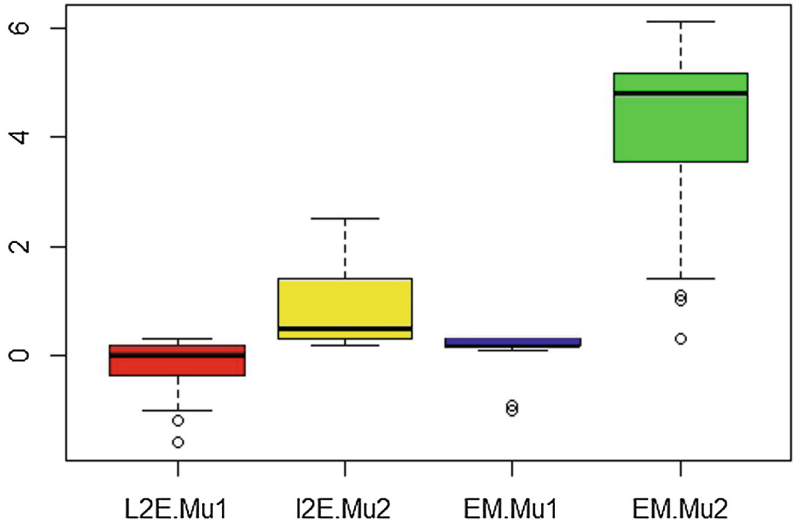


**Fig. 2.** Boxplots of the estimated mean for $L_2E$ and EM from 100 Monte Carlo Simulations of a GMM Model With Outliers

The boxplots of the parameter estimates of the component means for the mixture model in Eq. (8) with 1 % data contamination are shown in Fig. 2. The results clearly show that the outliers have a great influence on the EM method and that the $L_2E$ method is inherently robust to outliers.

### 4.2 Performance Due to Data/Model Mismatch

In this section, data/model mismatch is assessed. The robustness of $L_2E$ and EM is investigated when the postulated model is a mixture of Gaussians (GMM) but the data are generated from a mixture with symmetric departure from component normality. The setup as described in [12,18] is considered for the parameter estimation. More specifically, for the simulation study, a mixture with two components given by

$$f_{\boldsymbol{\theta}_2}(x) = \pi f_1(x) + (1 - \pi) f_2(x), \tag{9}$$

is considered. Note that $f_1$ is the density associated with a random variable $X_1 = aY_1$ ($a = 1$ chosen for the simulation) and $Y_1$ is a Student's $t(df)$-random

variable with a degree of freedom $df = 1$. Also, $f_2$ is the density associated with a random variable $X_2 = Y_2 + b$ ($b = 2$ chosen for the simulation) and $Y_2$ is a Student's $t(df)$-random variable with degrees of freedom $df = 4$. A total of 100 data points were generated and 50 Monte Carlo simulations were conducted to evaluate the performance of $L_2E$ and EM for consistency and efficiency by calculating the Bias and Mean Square Error (MSE).

Suppose $T(X)$ is an estimate of $\theta$. The Bias and MSE of $T$ are defined as

$$Bias(\theta) = E_\theta T - \theta \tag{10}$$

$$MSE(\theta) = E_\theta(T - \theta)^2 = Var_\theta(T) + Bias^2(\theta) \tag{11}$$

Note that the general shapes of such a two-component postulated (Gaussian mixture) model and a two-component $t$-mixture model from which the data are generated are different and further, the component densities in the sampling model have a much heavier tail than those in the postulated (Gaussian) mixture model. Table 1 depicts the bias and the mean square error for the mean estimates provided by the $L_2(E)$ and EM algorithms. The results show that the $L_2E$ is more robust than the EM approach with respect to data/model mismatch.

**Table 1.** Simulation results for data/model mismatch

| Estimation method | Component 1 | | Component 2 | |
|---|---|---|---|---|
| | Bias | MSE | Bias | MSE |
| $L_2E$ | 0.4 | 1.57 | 0.11 | 0.84 |
| $EM$ | −0.43 | 9.66 | 1.15 | 16.55 |

## 5 Summary and Conclusions

The $L_2E$ estimation technique can be easily constructed and applied to GMM and is a viable alternative to EM. Simulation studies revealed that the $L_2E$ mean estimates are robust to both outliers and data/model mismatch. The competitive performance of $L_2E$ make it stand out as an attractive alternative to EM for practical applications.

**Acknowledgment.** This work was supported by the National Science Foundation through Grant DUE-1122296.

## References

1. Titterington, D.M., Smith, A.F.M., Markov, U.E.: Statistical Analysis of Finite Mixture Distributions. Wiley, New York (1985)
2. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)

3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum-likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser. B **39**, 1–38 (1977)
4. Aitkin, M., Wilson, G.T.: Mixture models, outliers, and the EM algorithm. Technometrics **22**, 325–331 (1980)
5. Scott, D.W.: On fitting and adapting of density estimates. Comput. Sci. Stat. **30**, 124–133 (1998). (Weisberg, S., ed.)
6. Scott, D.W.: Remarks on fitting and interpreting mixture models. Comput. Sci. Stat. **31**, 104–109 (1999). (Berk, K., Pourahmadi, M., eds.)
7. Scott, D.W.: Parametric statistical modeling by minimum integrated square error. Technometrics **43**, 274–285 (2001)
8. Scott, D.W.: Outlier detection and clustering by partial mixture modeling. In: COMPSTAT Symposium. Physica-Verlag/Springer (2004)
9. Basu, A., Harris, I.R., Hjort, H.L., Jones, M.C.: Robust and efficient estimation by minimizing a density power divergence. Biometrika **85**, 549–560 (1998)
10. Markatou, M., Basu, A., Lindsay, B.G.: Weighted likelihood estimating equations with a bootstrap root search. J. Am. Stat. Assoc. **93**, 740–750 (1998)
11. Thayasivam, U., Sriram, T.N.: $L_2E$ estimation for mixture complexity for count data. Comput. Stat. Data Anal. **53**, 4243–4254 (2009)
12. Thayasivam, U., Sriram, T.N., Lee, J.: Simultaneous robust estimation in finite mixtures: the continuous case. J. Indian Stat. Assoc. **50**, 277–295 (2012)
13. Thayasivam, U., Shetty, S., Kuruwita, C., Ramachandran, R.P.: Detection of anomalies in network traffic using L2E for accurate speaker recognition. In: 55th International Midwest Symposium on Circuits & Systems, Boise, pp. 884–887 (2012)
14. Kai, Y., Dang, X., Bart, H., Chen, Y.: Robust model-based learning via Spatial-EM algorithm. IEEE Trans. Knowl. Data Eng. **27**, 1670–1682 (2015)
15. Rudemo, M.: Empirical choice of histograms and kernel density estimators. Scand. J. Statist. **9**, 65–78 (1982)
16. Bowman, A.W.: An alternative method of cross-validation for the smoothing of density estimates. Biometrika **71**, 353–360 (1984)
17. R: A Language and Environment for Statistical Computing, R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria, (2011). http://www.R-project.org/
18. Woodward, W.A., Parr, W.C., Schucany, W.R., Lindsay, H.: A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. J. Am. Stat. Assoc. **79**, 590–598 (1984)