MGMM: an R package for fitting Gaussian Mixture Models on Incomplete Genomics Data

Zachary R. McCaw [1,†,*], Hanna Julienne [2,†,*] and Hugues Aschard [2,3]

Abstract

Motivation: Although missing data are prevalent in genetic and genomics data sets, existing implementations of Gaussian mixture models (GMMs) require complete data. Standard practice is to perform complete case analysis or imputation prior to model fitting. Both approaches have serious drawbacks, potentially resulting in biased and unstable parameter estimates.
Results: Here we present `MGMM`, an `R` package for fitting GMMs in the presence of missing data. Using three case studies on real and simulated data sets, we demonstrate that MGMM is more effective at recovering true cluster assignments than standard GMM following state of the art imputation.
Availability: `MGMM` is available as an `R` package on Github at: https://github.com/zrmacc/MGMM.
Contact: zmccaw@alumni.harvard.edu, hanna.julienne@pasteur.fr

1 Introduction

Gaussian mixture models (GMMs) provide a flexible approach to multivariate density estimation and probabilistic clustering. Existing implementations of GMMs in the `R` computing environment, such as `mclust` and `mixtools`, require complete data. However, missing data are common in applications. Although widely applied, standard approaches for addressing missing data prior to clustering, including complete case analysis and imputation, have serious drawbacks. Complete case analysis makes inefficient use of the data, discarding information from observations that are only partially observed, which which leads to increased estimation variance. On the other hand, mean imputation introduces bias by making the incomplete observations appear less variable, and by shrinking the incomplete observations

† These authors contributed equally to this work. [1] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston MA, 02215, United States. [2] Groupe de Génétique Statistique, Département de Génomes and Génétique, C3BI, Institut Pasteur, Paris, France. [3] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, United States.

towards the complete data. This can result in inaccurate posterior membership probabilities that place excess weight on clusters with lower missingness. Although methods have been described for estimating GMMs from incomplete data (Ghahramani and Jordan, 1994), there are no existing implementations in `R`. To fill this gap, we present `MGMM`, an computationally efficient package for maximum likelihood (ML) estimation of GMMs in the presence of missing data. In contrast to complete case analysis, our approach makes full use of the available data, and in contrast to imputation, our approach is unbiased, and correctly propagates uncertainty. We present three case studies demonstrating that ML estimation via our expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993) outperforms imputation followed by standard GMM in recovering the true cluster assignments.

2 Methods

2.1 Setting

Consider n independent vectors $\boldsymbol{y}_i = \text{vec}(Y_{i1}, \cdots, Y_{ip})$ in $\mathbb{R}^p$, each arising from one of k distinct clusters. Let $Z_{ij} = 1$ if the ith observation belongs to the jth cluster, and define the $k \times 1$ indicator vector $\boldsymbol{z}_i = \text{vec}(Z_{i1}, \cdots, Z_{ik})$. Conditional on membership to the jth cluster, $\boldsymbol{y}_i$ follows a multivariate normal distribution, with cluster-specific mean $\boldsymbol{\mu}_j$ and covariance $\boldsymbol{\Sigma}_j$. Let π_j denote the marginal membership probability of the jth cluster. The observations can be viewed as arising from the following hierarchical model:

$$\begin{aligned} \boldsymbol{z}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}), \\ \boldsymbol{y}_i \big| (Z_{ij} = 1) &\sim N\big(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\big). \end{aligned} \tag{1}$$

Marginalized over the cluster assignment vector $\boldsymbol{z}_i$, each observation $\boldsymbol{y}_i$ follows a k component Gaussian Mixture Model (GMM):

$$f(\boldsymbol{y}_i) = \sum_{j=1}^{k} \pi_j f\big(\boldsymbol{y}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\big).$$

To perform estimation in the presence of missingness, we derive the EM objective function $Q(\boldsymbol{\pi}, \boldsymbol{\theta} | \boldsymbol{\pi}^{(r)}, \boldsymbol{\theta}^{(r)})$, which is the expectation of the complete data log likelihood, given the observed data and current parameter estimates. The EM objective is optimized using a sequence of three conditional maximizations. Let $\hat{\gamma}_{ij}^{(r)}$ denote the *responsibility* of the jth cluster for the ith observation, which is the current conditional probability of membership to that cluster, given the observed data. In the first step, the cluster means are updated using the responsibility-weighted average of the working outcome vectors $\hat{\boldsymbol{y}}_{ij}^{(r)}$. In the next step, the cluster covariances are updated using the responsibility-weighted average of the working residual outer products. In the final step, the cluster responsibilities and marginal membership probabilities are updated using the new means and covariances. This process iterates until the improvement in the EM objective drops below some tolerance ϵ (see supplementary notes).

2.2 Benchmarking against Imputation-based Methods

We designed a benchmarking procedure (see supplementary notes) to compare the performance of `MGMM` with imputation followed by standard GMM. The imputation methods included: naive mean and median imputation, k-nearest neighbors (kNN) imputation (Kowarik and Templ, 2016), multiple imputation by chained equations (MICE) (Buuren and Groothuis-Oudshoorn, 2010), and random forest imputation (Stekhoven and Bühlmann, 2011). We assessed the quality of the clustering by calculating the Adjust Rand Index (ARI) between the estimated and reference clusterings. An advantage of MGMM and MICE is that both provide an indication of the uncertainty in the cluster assignment. For each, we created an additional clustering rule in which observations with high assignment uncertainty were discarded (see supplementary notes).

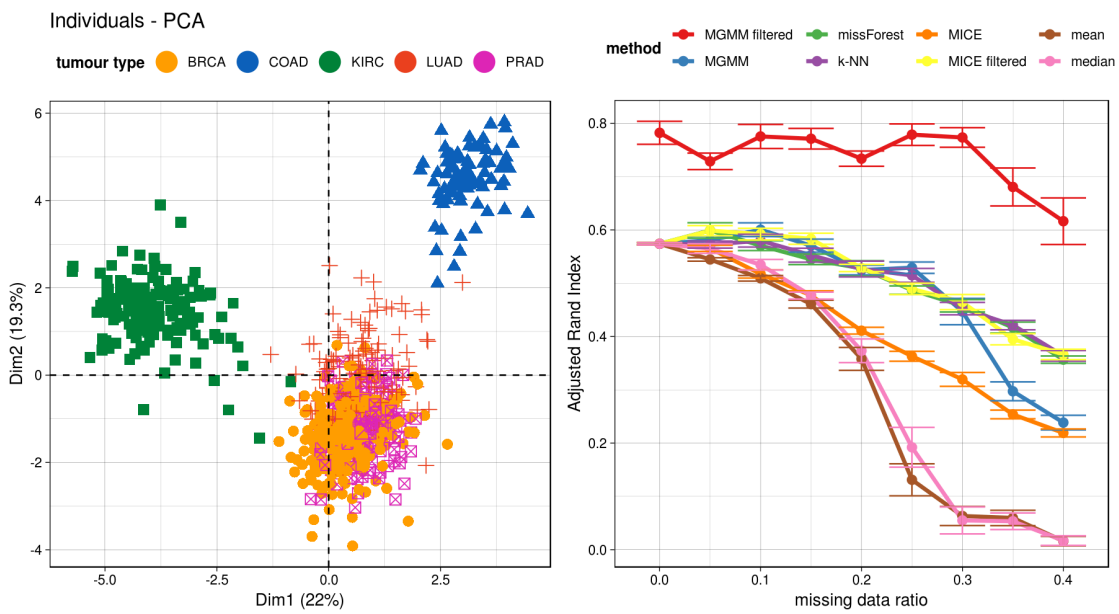

Figure 1: **Benchmarking for the Cancer RNA-Seq Data Set**. The left panel includes the projection of the expression data for $n = 801$ cancer patients onto the first two principal components. Observations are colored according to tumor type. The right panel presents the Adjusted Rand Index (ARI), comparing the estimated and reference clustering assignments, as a function of missingness level m. Error bars represent the standard error of the mean across 20 simulation replicates.

3 Results

We applied the benchmarking procedure to three case studies: a four component mixture of bivariate Gaussians, a cancer patient RNA-seq data set, and GWAS summary statistics for cardiovascular disease risk (Julienne et al., 2019).

3.1 Simulated Mixture of Gaussians Data

When the underlying distribution was a GMM, MGMM uniformly dominated imputation plus GMM (figure S1). Filtering out observations with elevated posterior entropy strikingly improve assignment accuracy. This suggest that MGMM correctly assess cluster assignment uncertainty, providing users with a mechanism for identifying observations with low confidence cluster assignments. Imputation by kNN and MICE performed similarly. Filtering out observations based on cluster assignment uncertainty only slightly improved the performance of MICE. Naive mean and median imputation strongly under-performed, and at elevated missingness created singularities in the data set that prevented the convergence of GMM.

3.2 Cancer Patient RNA-seq Data Set

To compare MGMM with imputation plus GMM in more realistic settings, we applied both to a Cancer RNA-Seq data set. Gene expression was measured in $n = 801$ patients having 1 of 5 tumor types. MGMM remained highly effective across missingness levels, especially when combined with entropy-based filtering. k-NN, MICE plus filtering, and missForest were competitive with base MGMM. MICE without filtering under-performed on these data, perhaps due to the presence of non-linear cluster boundaries. Mean and median imputation were not competitive at missingness exceeding 20%.

3.3 GWAS Summary Statistics

Finally, we considered clustering vectors of GWAS summary statistics arising when the same set of SNPs is tested for association with multiple traits. We considered a simulated data set, and three real data examples. For both simulated and real 3-trait data, the effect size heat map demonstrates clear cluster differentiation. MGMM with filtering performs competitively with the non-linear imputation strategies, kNN and missForest (see figures S3-S4). In the first 5-trait example (S5), cluster are clearly delineated and MGMM performs similarly to non linear imputation methods. The second 5-trait example exhibits poorly-separated clusters, which caused MGMM to under-perform at elevated missingness. This deterioration is reasonable given the drastic departure of the clusters from the assumed generative model.

4 Conclusion

We have presented `MGMM`, a powerful, general purpose `R` package for ML-based estimation of GMMs in the presence of missing data, and demonstrated that MGMM outperforms imputation plus standard GMM on various real and simulated data sets.

References

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68.

Ghahramani, Z. and Jordan, M. (1994). Supervised learning from incomplete data via an em approach. In *Advances in Neural Information Processing Systems 6*, pages 120–127. Morgan-Kaufmann.

Julienne, H., Lechat, P., Guillemot, V., Lasry, C., Yao, C., Laville, V., Vilhjalmsson, B., Ménager, H., and Aschard, H. (2019). Jass: Command line and web interface for the joint analysis of gwas results. *bioRxiv*.

Kowarik, A. and Templ, M. (2016). Imputation with the r package vim. *Journal of Statistical Software*, 74(7):1–16.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278.

Stekhoven, D. J. and Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.