Missingness Aware Gaussian Mixture Models

Zachary R. McCaw

2020-07-26

Contents

Introduction

Overview

This package implements clustering of multivariate normal random vectors with missing elements. Clustering is achieved by fitting a Gaussian Mixture Model (GMM). The parameters are estimated by maximum likelihood, using the Expectation Maximization (EM) algorithm. Our implementation extends existing methods by allowing for missingness in the input data. The EM algorithm addresses missingness of both the cluster assignments and the vector components. The output includes the marginal cluster membership probabilities; the mean and covariance of each cluster; the density of each mixture component evaluated at the observations; the posterior probabilities of cluster membership; maximum *a posteriori* cluster assignments; and a completed version of the input data, with missing values replaced by their posterior expectations.

Model

Suppose the data consist of n random vectors in $\mathbb{R}^D$. Each observation Y_i arises from one of K distinct clusters. Associate with each observation a $K \times 1$ vector of indicators $Z_i \in \{0,1\}$, where $Z_{ik} = 1$ if observation i belongs to cluster k, and $Z_{ik} = 0$ otherwise. These cluster membership indicators are latent variables. The marginal probability of membership to cluster k is $\pi_k = P(z_{ik} = 1)$. Conditional on membership to the kth cluster, the observation Y_i follows a multivariate normal distribution, with cluster-specific mean μ_k and covariance Σ_k.Overall, the generative model is:

$$Z_i \sim \text{Multinomial}(\pi_1, \cdots, \pi_K)$$

$$Y_i \big| (Z_{ik} = 1) \sim N\big(\mu_k, \Sigma_k\big)$$

Marginally, the observations Y_i follow a GMM, with density:

$$f(y_i) = \sum_{k=1}^{K} f(y_i, z_{ik} = 1) = \sum_{k=1}^{K} f(y_i | z_{ik} = 1) P(z_{ik} = 1) = \sum_{k=1}^{K} f(y_i | \mu_k, \Sigma_k) \pi_k.$$

Each element Y_{id} of Y_i is potentially missing at random. Associated with each observation a $D \times 1$ vector of response indicators R_i, where $R_{id} = 1$ if element d of observation i is observed, and $R_{id} = 0$ otherwise. Partition each Y_i into its observed Y_i^{obs} and missing Y_i^{mis} components. That is, element Y_{id} belongs to Y_i^{obs} if $R_{id} = 1$, and belongs to Y_i^{mis} if $R_{id} = 0$. The missingness occurs at random if $(R_{id} \perp Y_{id})|Y_i^{\text{obs}}$. That is, given the observed elements of Y_i, whether any remaining element is missing is independent of that element's value.

Maximum likelihood estimates (MLEs) for the parameters of the GMM are obtained using the EM algorithm. During the E-step, both the cluster assignments Z_{ik} and the unobserved components Y_i^{mis} of Y_i are treated as missing data. Suppose momentarily that all data were observed for observation i. The contribution of subject i to the *complete data* log likelihood would be:

$$\ell_i = \sum_{k=1}^{K} z_{ik} \ln \pi_k - \frac{1}{2} \sum_{k=1}^{K} z_{ik} \ln \det(\Sigma_k) - \frac{1}{2} \sum_{k=1}^{K} z_{ik} (y_i - \mu_k) \Sigma_k^{-1} (y_i - \mu_k).$$

Since the Z_{ik} are not observed, and the Y_i are incompletely observed, the complete data log likelihood cannot be evaluated. Instead, an EM objective is formed by taking the expectation of the complete data log likelihood, conditional on the observed data and the current parameter state:

$$q_i^{(r)} \equiv E\big\{\ell_i | y_i^{\text{obs}}, \vartheta^{(r)}\big\} = \sum_{k=1}^{K} \gamma_{ik}^{(r)} \ln \pi_k - \frac{1}{2} \sum_{k=1}^{K} \gamma_{ik}^{(r)} \ln \det(\Sigma_k) - \frac{1}{2} \sum_{k=1}^{K} \operatorname{tr}\big\{\Sigma_k^{-1} V_{ik}^{(r)}\big\}.$$

Here y_i^{obs} is the observed data for subject i; $\vartheta^{(r)}$ is the current parameter state; $\gamma_{ik}^{(r)}$ is the *responsibility*, defined as $E(Z_{ik}|y_i^{\text{obs}}, \vartheta^{(r)}]$; and $V_{ik}^{(r)}$ is the *expected residual outer product*, defined as $E\big\{Z_{ik}(Y_i - \mu_k)(Y_i - \mu_k)'|y_i^{\text{obs}}, \vartheta^{(r)}\big\}$. In the M-step, updates of the model parameters ϑ are obtained by maximizing the EM objective.

Once the convergence criterion has been achieved, the responsibility, or posterior probability of cluster membership, is calculated as:

$$\gamma_{ik} = P(Z_{ik} = 1|y_i^{\text{obs}}) = \frac{f(y_i^{\text{obs}}|\ \mu_k, \Sigma_k)\pi_k}{\sum_{k'=1}^{K} f(y_i^{\text{obs}}|\ \mu_{k'}, \Sigma_{k'})\pi_{k'}}.$$

The maximum a posteriori classification for y_i is given by:

$$A_i = \arg\max_k\ \gamma_{ik}$$

For observation Y_i, posterior expectation of the missing elements Y_i^{mis}, given the observed elements Y_i^{obs}, is:

$$E(Y_i^{\text{mis}}|y_i^{\text{obs}}) = \sum_{k=1}^{K} E(Y_i^{\text{mis}}|y_i^{\text{obs}}, z_{ik} = 1)\pi_k.$$

Data Generation

Description

The function `rGMM` simulates observations from a Gaussian Mixture Model. The number of observations is specified by `n`, and the dimension of each observation by `d`. The number of clusters is set using `k`, which defaults to 1. The marginal probabilities of cluster membership are provided as a numeric vector `pi`, which should contain `k` elements. If `pi` is omitted, the clusters are assumed equi-probable. The proportion of elements in the $n \times d$ data matrix that are missing is specified by `miss`, which defaults to zero. Note that

when `miss>0`, it is possible for all elements of an observation to be missing. The cluster `means` are provided either as a numeric prototype vector, or as a list of numeric vectors. If a single prototype is provided, that vector is taken as the mean for all clusters. By default, the zero vector is adopted as the prototype. The cluster covariances `covs` are provided as a numeric prototype matrix, or as a list of such matrices. If a single prototype is provided, that matrix is used as the covariance for all clusters. By default, the identity matrix is adopted as the prototype.

Examples

Single Component without Missingness

In this example, `n = 1e3` observations are simulated from a single `k = 1` bivariate normal distribution `d = 2` without missingness. The mean is $\mu = (2, 2)$, and the covariance is an exchangeable correlation structure with off-diagonal $\rho = 0.5$.

```
set.seed(100)
# Single component without missingness
cov <- matrix(c(1, 0.5, 0.5, 1), nrow = 2)
data <- rGMM(n = 1e3, d = 2, k = 1, means = c(2, 2), covs = cov)
```

Single Component with Missingness

In this example, `n = 1e3` observations are simulated from a single `k = 1` trivariate normal distribution `d = 3` with 20% missingness `miss = 0.2`. The mean defaults to the zero vector, and the covariance to the identity matrix.

```
# Single component with missingness
data <- rGMM(n = 1e3, d = 3, k = 1, miss = 0.2)
```

Two Components without Missingness

In this example, `n = 1e3` observations are simulated from a two-component `k = 2` trivariate normal distribution `d = 3` without missingness. The mean vectors are $\mu_1 = (-2, -2, -2)$ and $\mu_2 = (2, 2, 2)$. The covariance matrices are both exchangeable with off-diagonal $\rho = 0.5$. Since `pi` is omitted, the cluster are equi-probable, i.e. $\pi_1 = \pi_2 = 1/2$.

```
# Two-component mixture without missingness
mean_list <- list(
  c(-2, -2, -2),
  c(2, 2, 2)
)
cov <- matrix(
  c(1, 0.5, 0.5,
    0.5, 1, 0.5,
    0.5, 0.5, 1), nrow = 3)
data <- rGMM(n = 1e3, d = 3, k = 2, means = mean_list, covs = cov)
```

Four Components with Missingness

In this example, `n = 1e3` observations are simulated from a four-component `k = 4` bivariate normal distribution `d = 2` with 10% missingness `miss = 0.1`. The mean vectors are $\mu_1 = (-2, -2)$, $\mu_2 = (-2, 2)$,

$\mu_3 = (2, -2)$ and $\mu_4 = (2, 2)$. The covariance matrices are all $0.5 * I$. The cluster proportions are (35%, 15%, 15%, 35%) for $(\pi_1, \pi_2, \pi_3, \pi_4)$, respectively.

```
# Four-component mixture with missingness
mean_list <- list(
  c(-2, -2),
  c(-2, 2),
  c(2, -2),
  c(2, 2)
)
cov <- 0.5 * diag(2)
props <- c(0.35, 0.15, 0.15, 0.35)
data <- rGMM(
  n = 1e3,
  d = 2,
  k = 4,
  pi = props,
  miss = 0.1,
  means = mean_list,
  covs = cov
)
```

Parameter Estimation

Description

The function `fit.GMM` estimates the GMM parameters. The data are expected as a numeric matrix `data`, with observations as rows. The number of mixture components is specified using `k`, which defaults to 1. Initial values for the mean vectors, covariance matrices, and cluster proportions are provided using `init_means`, `init_covs`, and `init_props`, respectively. The initial means `init_means` are provided as a list of vectors, the initial covariances `init_covs` as a list of matrices, and the cluster proportions `init_props` as a numeric vector. Initial means and covariances should be supplied as a lists with `k` components, even if `k = 1`, or if all `k > 1` components are initialized at the same value.

If the `data` contain complete observations, i.e. observations with no missing elements, `fit.GMM` will attempt to initialize all model parameters (μ, Σ, π). However, if the data `data` contain no complete observations, then initial values are required for each of `init_means`, `init_props`, and `init_props`. Supplying initial values may also result in better performance when there are relatively few complete observations.

The arguments `maxit`, `eps`, and `report` control the fitting procedure. `maxit` sets the maximum number of EM iterations to attempt. The default is 10^2. `eps` sets the minimum acceptable improvement in the EM objective function. The default is 10^{-6}. If `report = TRUE`, then fitting progress is displayed.

Examples

Single Component without Missingness

In this example, 10^3 observations are simulated with a single bivariate normal distribution without missingness. Since the model contains only a single component, the output is a list containing the estimated mean and covariance, and the model log likelihood. In the case of a single component without missingness, the maximum likelihood estimates are available in closed form.

```r
# Single component without missingness
cov <- matrix(c(1, 0.5, 0.5, 1), nrow = 2)
data <- rGMM(n = 1e3, d = 2, k = 1, means = c(2, 2), covs = cov)
fit <- fit.GMM(data, k = 1)
cat("Estimated Mean and Covariance:\n")
show(fit)
```

```
## Estimated Mean and Covariance:
## $Mean
##       y1       y2
## 2.024176 2.007562
##
## $Covariance
##           [,1]      [,2]
## [1,] 1.0669079 0.5317859
## [2,] 0.5317859 0.9994667
##
## $Objective
## [1] -1754.07
```

Single Component with Missingness

In this example, 10^3 observations are simulated from a single trivariate normal distribution with 20% missingness. Since the model contains only a single component, the output is again a list. However, in the case of missingness, the EM algorithm is used for estimation, and a completed version of the input data is returned, with missing values replaced by their posterior expectations.The true mean is the zero vector, and the true covariance is identity. For `fit1` below, the initial mean and covariance are estimated internally using complete observations. For `fit2` below, the mean and covariance are initialized at the truth. The final value of the EM objective is increased by initializing at the truth.

```r
set.seed(102)

# Single component with missingness
data <- rGMM(n = 1e3, d = 3, k = 1, miss = 0.2)

cat("Initial parameter values set internally:\n")
fit1 <- fit.GMM(data, k = 1)

cat("\nEstimated mean:\n")
show(fit1$Mean)

cat("\nEstimated covariance:\n")
show(fit1$Covariance)

cat("\nFinal objective:\n")
show(fit1$Objective)

cat("Initial parameter values set manually:\n")
init_means = list(
  c(0, 0, 0)
)
init_covs = list(
  diag(3)
)
```

```r
fit2 <- fit.GMM(data, k = 1, init_means = init_means, init_covs = init_covs)

cat("\nEstimated mean:\n")
show(fit2$Mean)

cat("\nEstimated covariance:\n")
show(fit2$Covariance)

cat("\nFinal objective:\n")
show(fit2$Objective)

cat("\nGain in final objective by initializing parameters at the truth:\n")
fit2$Objective-fit1$Objective
```

```
## Initial parameter values set internally:
## Objective increment:  1.32
## Objective increment:  0.0425
## Objective increment:  0.00177
## Objective increment:  0.000115
## Objective increment:  1.13e-05
## Objective increment:  1.34e-06
## Objective increment:  1.68e-07
## 6 update(s) performed before reaching tolerance limit.
##
##
## Estimated mean:
##           y1           y2           y3
##  0.007209316  0.053596831 -0.027423821
##
## Estimated covariance:
##             y1          y2         y3
## y1  0.91340883 -0.02162862 0.02771232
## y2 -0.02162862  0.97258027 0.06271627
## y3  0.02771232  0.06271627 0.95024429
##
## Final objective:
## [1] -2793.743
## Initial parameter values set manually:
## Objective increment:  8.47
## Objective increment:  0.541
## Objective increment:  0.0465
## Objective increment:  0.00493
## Objective increment:  0.000572
## Objective increment:  6.87e-05
## Objective increment:  8.37e-06
## Objective increment:  1.03e-06
## Objective increment:  1.26e-07
## 8 update(s) performed before reaching tolerance limit.
##
##
## Estimated mean:
##           y1           y2           y3
##  0.007209512  0.053596653 -0.027423582
##
```

```
## Estimated covariance:
##             y1          y2         y3
## y1  0.91340891 -0.02163085 0.02771294
## y2 -0.02163085  0.97258018 0.06271506
## y3  0.02771294  0.06271506 0.95024417
##
## Final objective:
## [1] -2793.743
##
## Gain in final objective by initializing parameters at the truth:
## [1] 0.0001295088
```

Two Components without Missingness

In this example, 10^3 observations are simulated from a two-component, trivariate normal distribution without missingness. Since the model has multiple components, the output is an object of class `mix`. The *show method* displays the estimated cluster proportions and the final objective. The slots of the output contain the following: * `@Means` and `@Covariances`: lists of the estimated cluster means and covariances. * `@Density`: the cluster densities evaluated at the observations. * `@Responsibilities`: the posterior membership probabilities for each observation. * `@Assignments`: the maximum a posteriori cluster assignments and assignment entropy. * `@Completed`: a completed version of the input data is returned, with missing values replaced by their posterior expectations

```r
# Two componets without missingness
mean_list <- list(
  c(-2, -2, -2),
  c(2, 2, 2)
)
cov <- matrix(
  c(1, 0.5, 0.5,
    0.5, 1, 0.5,
    0.5, 0.5, 1), nrow = 3
)

data <- rGMM(n = 1e3, d = 3, k = 2, means = mean_list, covs = cov)
fit <- fit.GMM(data, k = 2, maxit = 10, eps = 1e-8)

cat("\n")
show(fit)

cat("Cluster means:\n")
fit@Means

cat("Cluster covariances:\n")
fit@Covariances

cat("Cluster responsibilities:\n")
head(fit@Responsibilities)

cat("\nCluster assignments:\n")
head(fit@Assignments)

cat("\nCompleted data:\n")
head(fit@Completed)
```

```
## Objective increment:  0.627
## Objective increment:  0.0673
## Objective increment:  0.0175
## Objective increment:  0.00462
## Objective increment:  0.00124
## Objective increment:  0.000331
## Objective increment:  8.86e-05
## Objective increment:  2.37e-05
## Objective increment:  6.37e-06
## Objective increment:  1.71e-06
## 10 update(s) performed without reaching tolerance limit.
##
##
## Normal Mixture Model with 2 Components.
## Cluster Proportions:
##    k1    k2
## 0.514 0.486
##
## Final Objective:
## [1] -3019.68
##
## Cluster means:
## [[1]]
##       y1       y2       y3
## 1.993715 2.044410 2.042874
##
## [[2]]
##        y1        y2        y3
## -2.023218 -1.967009 -2.031333
##
## Cluster covariances:
## [[1]]
##           y1        y2        y3
## y1 1.0421717 0.5260849 0.5547371
## y2 0.5260849 1.0938308 0.5580000
## y3 0.5547371 0.5580000 0.9795204
##
## [[2]]
##           y1        y2        y3
## y1 0.9486886 0.4428439 0.4698537
## y2 0.4428439 0.9356663 0.4986851
## y3 0.4698537 0.4986851 1.0774598
##
## Cluster responsibilities:
##             k1           k2
## 1 7.213209e-06 9.999928e-01
## 2 9.999999e-01 1.338099e-07
## 2 9.999770e-01 2.298827e-05
## 2 8.135063e-02 9.186494e-01
## 1 5.636022e-06 9.999944e-01
## 2 1.000000e+00 3.279117e-08
##
## Cluster assignments:
##   Assignments      Entropy
```

```
## 1           2 1.336147e-04
## 2           1 3.248369e-06
## 2           1 3.873851e-04
## 2           2 4.069204e-01
## 1           2 1.064057e-04
## 2           1 8.625655e-07
##
## Completed data:
##           y1         y2         y3
## 1 -3.5781991 -1.7243246 -0.7998434
## 2  2.4957664  3.1465942  1.6865892
## 2  1.4038862  1.7525323  2.0079278
## 2 -0.8862729 -1.2629163  1.0069399
## 1 -1.8648070 -0.7268469 -3.3138250
## 2  2.4270751  2.6503247  3.3227434
```

Four Components with Missingness

In this example, 10^3 observations are simulated from a four-component bivariate normal distribution with 10% missingness. Since the model has multiple components, the output is an object of class `mix`.

```r
set.seed(200)

# Four components with missingness
mean_list <- list(
  c(2, 2),
  c(2, -2),
  c(-2, 2),
  c(-2, -2)
)
cov <- 0.5 * diag(2)
props <- c(0.35, 0.15, 0.15, 0.35)
data <- rGMM(
  n = 1000,
  d = 2,
  k = 4,
  pi = props,
  miss = 0.1,
  means = mean_list,
  cov = cov
)
fit <- fit.GMM(data, k = 4, maxit = 10, eps = 1e-8)
show(fit)

cat("Cluster means:\n")
fit@Means

cat("Cluster covariances:\n")
fit@Covariances

cat("\nCluster assignments:\n")
head(fit@Assignments)
```

```
## Objective increment:  1.57
```

```
## Objective increment:  0.161
## Objective increment:  0.0334
## Objective increment:  0.00726
## Objective increment:  0.0016
## Objective increment:  0.000341
## Objective increment:  7.46e-05
## Objective increment:  1.59e-05
## Objective increment:  3.48e-06
## Objective increment:  7.47e-07
## 10 update(s) performed without reaching tolerance limit.
##
## Normal Mixture Model with 4 Components.
## Cluster Proportions:
##    k1    k2    k3    k4
## 0.149 0.388 0.130 0.332
##
## Final Objective:
## [1] -1888.25
##
## Cluster means:
## [[1]]
##        y1        y2
## -2.046472  2.074567
##
## [[2]]
##        y1        y2
## -2.026168 -1.983955
##
## [[3]]
##        y1        y2
##  2.026073 -2.075973
##
## [[4]]
##       y1       y2
## 1.973530 1.981959
##
## Cluster covariances:
## [[1]]
##            y1         y2
## y1 0.43656741 0.02040745
## y2 0.02040745 0.45492340
##
## [[2]]
##             y1          y2
## y1 0.567726041 0.003079349
## y2 0.003079349 0.575620933
##
## [[3]]
##             y1          y2
## y1  0.39075316 -0.06937548
## y2 -0.06937548  0.37947376
##
## [[4]]
##             y1          y2
```

```
## y1 0.484058646 0.002193949
## y2 0.002193949 0.525919652
##
##
## Cluster assignments:
##   Assignments      Entropy
## 1           4 2.226854e-10
## 4           2 4.292356e-01
## 2           3 6.473445e-02
## 2           3 1.011832e-07
## 1           4 2.032369e-08
## 4           2 3.292494e-04
```

Cluster Number Selection

Clustering Quality

The function `ClustQual` provides several metrics for internally assessing the quality of cluster assignments from a fitted GMM. The input is an object of class `mix`. The output is a list containing the metrics: BIC, CHI, DBI, and SIL.

- BIC is the Bayesian Information Criterion, which is a penalized version of the negative log likelihood. A lower value indicates better clustering quality.
- CHI is the Calinski-Harabaz Index, a ratio of the between cluster to within cluster variation. A higher value indicates better clustering quality.
- DBI is the Davies-Bouldin Index, an average of cluster similarities. A lower value indicates better clustering quality.
- SIL is the average Silhouette width, a measure of how well an observation matches its assigned cluster. A higher value indicates better clustering quality.

```r
set.seed(105)

# Four components without missingness
mean_list <- list(
  c(2, 2),
  c(2, -2),
  c(-2, 2),
  c(-2, -2)
)
cov <- 0.5 * diag(2)
data <- rGMM(n = 100, d = 2, k = 4, means = mean_list)
fit <- fit.GMM(data, k = 4, maxit = 100, eps = 1e-8, report = F)

# Quality metrics
clust_qual = ClustQual(fit)

cat("BIC:\n")
clust_qual$BIC

cat("\nCHI:\n")
clust_qual$CHI
```

```r
cat("\nDBI:\n")
clust_qual$DBI

cat("\nSIL:\n")
clust_qual$SIL
```

```
## BIC:
## [1] 358.6248
##
## CHI:
## [1] 6.546848
##
## DBI:
## [1] 0.5617234
##
## SIL:
## [1] 0.545614
```

Choosing the Number of Clusters

In applications, the number of clusters k is often unknown. The function `ChooseK` is designed to assist in choosing the number of clusters. The inputs include the data matrix `data`, the minimum cluster number to assess `k0`, the maximum cluster number to assess `k1`, and the number of bootstrap replicates at each cluster number `boot`. For each cluster number k between k_0 and k_1, `boot` bootstrap data sets are generated. A GMM with k components is fit, and the quality metrics are calculated. The bootstrap replicates are summarized by their mean and standard error (SE). For each quality metric, the cluster number k_{opt} that had the optimal quality, and the smallest cluster number whose quality was within 1 SE of the optimum $k_{1\text{se}}$, are reported. The output is a list `Choices` containing the cluster numbers selected by each metric, and the complete set of bootstrap `Results`.

```r
# Cluster number selection
choose_k = ChooseK(data, k0 = 2, k1 = 6, boot = 10)

cat("\nCluster number choices:\n")
choose_k$Choices

cat("\nAll results:\n")
head(choose_k$Results)
```

```
## Cluster size 2 complete. 11 fit(s) succeeded.
## Cluster size 3 complete. 11 fit(s) succeeded.
## Cluster size 4 complete. 11 fit(s) succeeded.
## Cluster size 5 complete. 11 fit(s) succeeded.
## Cluster size 6 complete. 11 fit(s) succeeded.
##
## Cluster number choices:
##   Metric k_opt  Metric_opt k_1se  Metric_1se
## 1    BIC     6 289.8829238     6 289.8829238
## 2    CHI     6  10.2082679     6  10.2082679
## 3    DBI     4   0.5584643     4   0.5584643
## 4    SIL     4   0.5657591     4   0.5657591
##
## All results:
##   Clusters Fits Metric        Mean          SE
```

```
## 1       2   11    BIC 449.9947433 5.233711804
## 2       2   11    CHI   1.6024639 0.039332235
## 3       2   11    DBI   1.0222610 0.012683919
## 4       2   11    SIL   0.4372601 0.004176608
## 5       3   11    BIC 369.7079292 9.428310875
## 6       3   11    CHI   3.4225779 0.111995937
```