High-Dimensional Gaussian Mixtures with Random Projection Based Covariance Estimates

High-Dimensional Gaussian Mixtures with Random Projection Based Covariance Estimates

Silvia Dallari*, Laura Anderlucci, and Angela Montanari

Department of Statistical Sciences, University of Bologna, Bologna, Italy
{silvia.dallari2, laura.anderlucci, angela.montanari}@unibo.it

Abstract. Random projections (RPs) are known to provide promising results in the context of high-dimensional supervised classification, but even when information on the class membership is not available, the classification issue can be addressed by exploiting the general idea of RP ensemble. In this work, we address the problem of clustering high-dimensional data in the Gaussian mixture model (GMM) framework by resorting to a RP-based ensemble of low-rank estimates for the group-specific covariance matrix. When the number of features is large compared to the number of units, the most widely used solution for GMM estimation employs parsimonious covariance parameterizations via spectral decomposition, in order to cope with possibly singular group-specific covariance matrices. Our approach offers an alternative to these parsimonious parameterizations which guarantees a full rank estimate for each covariance matrix, thus allowing for the estimation of mixtures of Gaussian graphical models, too. The potential of the proposal is demonstrated through a real data application.

Keywords: Random projections, Model-based clustering, high-dimensional covariance estimation

1 Introduction

Cluster analysis aims to partition the data into groups so that observations within the same cluster are more homogeneous to each other than to those in different clusters (for a comprehensive overview, see [5]).

While clustering in low-dimensional spaces is a relatively simple task, the complexity of the procedure increases as the number of variables, denoted as p, grows. Traditional unsupervised classification methods present several limitations when applied to high-dimensional data, due for example to the presence of noisy or irrelevant features.

Model-based clustering ([1],[9]) assumes that data come from a common source with k different sub-populations. Each sub-population, or cluster, is described by a specific density function and the overall population is a mixture of them. The resulting model is a finite mixture and, when the densities are multivariate Gaussians, it is defined by the following probability density function:

* Corresponding author.

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^{k} \pi_i \phi_i(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i),$$

where $\boldsymbol{\pi}$ is the vector of mixture components, that satisfy $\sum_{i=1}^{k} \pi_i = 1$ and $\pi_i > 0 \; \forall \; i$, and $\phi_i(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ is a multivariate Gaussian density function parameterized by its mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$:

$$\phi_i(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) = (2\pi)^{-(p/2)}|\boldsymbol{\Sigma}_i|^{-(1/2)} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}.$$

In order to deal with the high-dimensionality issue, several approaches have been proposed in the literature, from variable selection (e.g., [7] and [13]) to penalized models that encourage sparsity in the variables (e.g., [15]); from dimension reduction to parsimonious models (see, e.g., [1], [4], and [11]) and to subspace methods (e.g., [10], [3]). A thorough review of the various approaches can be found, among others, in [2].

Within the dimensionality reduction class, Random Projections (RPs) proved to be a useful tool able to uncover the underlying group structure.

In this paper, we propose a novel approach to cluster high-dimensional data in the Gaussian mixture model (GMM) framework by resorting to a RP-based ensemble of low-rank estimates for the group-specific covariance matrix.

2 Random Projections

Random projections have been extensively employed in all the fields of multivariate analysis in the past decades, from large covariance matrix estimation to supervised classification, from sparse PCA to multiple linear regression. The most appealing feature is their ability to reduce the dimensionality of a data matrix while preserving most of the information.

Specifically, given a high-dimensional matrix $\mathbf{X}$ of dimension $n \times p$, a lower dimensional random projection matrix $\mathbf{A}$ of dimension $p \times d$, with $d \ll p$, can be considered, such that the random linear combination of the columns of $\mathbf{X}$ with weights given by the elements of $\mathbf{A}$ produces a projected matrix $\mathbf{Y}$ of reduced dimension $n \times d$.

Methods for multivariate analysis based on random projections are usually employed according to the following steps:

1. Map at random the original high-dimensional data onto a lower subspace;
2. Apply the chosen method to the dimensionally reduced data;
3. Combine the results of the analysis on a set of selected Random Projections.

3 Gaussian Mixture Models with Random Projection Ensemble Covariance Estimate

A strategy to obtain a full-rank estimate that exploits random projections has been recently proposed by Marzetta et al. [6]. In their work, they estimate a full rank covariance matrix from a collection of low rank covariance matrices via ensembles of RPs. More in details, their method works as follows:

Step 1: Perform the analysis on the reduced space

1. Consider, without loss of generality, a mean-centered data matrix $\mathbf{X}$ of dimension $n \times p$ and a Haar projection matrix $\mathbf{A}$ of dimension $p \times d$ ($d \ll p$). Then, a non singular lower-dimensional $d \times d$ covariance matrix is:

$$\mathbf{S}^{(d)} = \frac{1}{n}\mathbf{Y}^\top\mathbf{Y} = \frac{1}{n}\mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{A} = \mathbf{A}^\top\mathbf{S}\mathbf{A},$$

   with $\mathbf{S} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$.
2. Project out the $\mathbf{S}^{(d)}$ covariance matrix to a $p \times p$ covariance matrix using the same projection matrix $\mathbf{A}$:

$$\mathbf{S}^{(p)} = \frac{1}{n}\mathbf{A}(\mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{A})\mathbf{A}^\top = \mathbf{A}\mathbf{S}^{(d)}\mathbf{A}^\top$$

Step 2: Ensemble

1. Consider a set B of random projections.
2. Aggregate the B covariance matrices $\mathbf{S}^{(p)}$ by taking the expected value over the ensemble:

$$_{RP}\hat{\boldsymbol{\Sigma}} = E_{\mathbf{A}}[\mathbf{A}\mathbf{S}^{(d)}\mathbf{A}^\top] = E_{\mathbf{A}}[\mathbf{A}(\mathbf{A}^\top\mathbf{S}\mathbf{A})\mathbf{A}^\top]$$

It is reasonable to rescale the above covariance expression by the factor p/d, because the dimensionality reduction leads to shortened feature vectors whose norm is generally d/p times that of the original ones [6].

The expected value of the ensemble can be asymptotically obtained in closed form using Schur polynomials.

We propose to extend the proposal of Marzetta et al. [6] to the Gaussian mixture model framework when bad conditioning issues are present. Specifically, their solution can be used to estimate the group-specific covariance matrices within the framework of a Gaussian mixture model. This approach provides an alternative to the parsimonious covariance structures present in the `mclust` library ([14]).

In order to speed computations, we have always employed the asymptotic approximation of the ensemble estimate. As the optimal value of d depends both on the number of units and on the number of variables, a thorough simulation study, aimed at providing rule of thumb criteria for the selection of d, has been performed.

4 Real Data Analysis

The method has been tested on both real and simulated data.

In this section, we analyze a dataset that consists of 231 samples and 1050 variables of homogenized raw meat from five different animal species (beef, chicken, lamb, pork and turkey). The data were collected across a wavelength range of 400 to 2498 nm, with measurements recorded every 2 nm ([8]). The information about the original data source and the conditions for the use of the data are outlined in the online supplementary material of [12].

We compared our proposal (GMM-RPCov) with $d = 10$ to the results of a GMM fit to the original uncompressed data. Adjusted Rand indexes (ARI) are reported in **Table 1**. For the group-specific covariance matrices, function `mclust` selects the `EVI` model, that is diagonal, with equal volume and varying shape.

Table 1. Adjusted Rand Indexes obtained for the meat dataset with GMM and GMM-RPCov ($d = 10$).

	GMM	GMM-RPCov
ARI	0.14	0.38

Although the resulting ARI is still relatively low, the confusion matrix in **Table 2** shows a reasonable partition. In fact, the model is able to recognise the different kinds of red meats, while it tends to group white meats together.

Table 2. Optimal classification obtained for the meat dataset.

	Beef	Chicken	Lamb	Pork	Turkey
1	0	0	33	0	0
2	0	3	0	21	1
3	0	30	0	5	31
4	0	22	0	29	23
5	32	0	1	0	0

The proposed approach to cluster high-dimensional data demonstrates to be effective in uncovering the underlying group structure; in addition, it is able to provide estimates of the group-specific covariance matrices that are unconstrained, in the `mclust` sense, and full rank.

Acknowledgments

This research received funding by MUR-PNRR M4C2I1.3 PE6 project PE00000019 HEAL ITALIA.

References

1. Banfield, J.D., Raftery, A.E.: Model-Based Gaussian and Non-Gaussian Clustering. Biometrics **49**(3), 803–821 (1993)
2. Bouveyron, C., Brunet-Saumard, C.: Model-based clustering of high-dimensional data: A review. Computational Statistics & Data Analysis **71**, 52–78 (2014)
3. Bouveyron, C., Girard, S., Schmid, C.: High-dimensional data clustering. Computational Statistics & Data Analysis **52**(1), 502–519 (2007)
4. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognition **28**(5), 781–793 (1995)
5. Hennig, C., Meila, M., Murtagh, F., Rocci, R.: Handbook of Cluster Analysis. CRC Press (2015)
6. Marzetta, T.L., Tucci, G.H., Simon, S.H.: A random matrix-theoretic approach to handling singular covariance estimates. IEEE Transactions on Information Theory **57**(9), 6256–6271 (2011)
7. Maugis, C., Celeux, G., Martin-Magniette, M.L.: Variable selection in model-based clustering: A general variable role modeling. Computational Statistics & Data Analysis **53**(11), 3872–3882 (2009)
8. McElhinney, J., Downey, G., Fearn, T.: Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. Journal of Near Infrared Spectroscopy **7**, 145–154 (1999)
9. McLachlan, G.J., Peel, D.: Finite Mixture Models. John Wiley & Sons, Inc. (2000)
10. McLachlan, G.J., Peel, D., Bean, R.W.: Modelling high-dimensional data by mixtures of factor analyzers. Computational Statistics & Data Analysis **41**(3-4), 379–388 (2003)
11. McNicholas, P.D., Murphy, T.B.: Parsimonious Gaussian mixture models. Statistics and Computing **18**, 285–296 (2008)
12. Murphy, T.B., Dean, N., Raftery, A.E.: Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. The Annals of Applied Statistics **4**(1), 396 – 421 (2010)
13. Raftery, A.E., Dean, N.: Variable selection for model-based clustering. Journal of the American Statistical Association **101**(473), 168–178 (2006)
14. Scrucca, L., Fraley, C., Murphy, T.B., Raftery, A.E.: Model-Based Clustering, Classification, and Density Estimation Using mclust in R. Chapman and Hall/CRC Press (2023)
15. Wang, S., Zhu, J.: Variable selection for model-based high-dimensional clustering and its application to microarray data. Biometrics **64**(2), 440–448 (2008)