Flexible clustering methods for high-dimensional data sets.

Metodi di cluster analysis flessibili per data set di grandi dimensioni

Cristina Tortora and Paul D. McNicholas

Abstract Finite mixture models assume that a population is a convex combination of densities; therefore, they are well suited for clustering applications. Each cluster is modeled using a density function. One of the most flexible distributions is the generalized hyperbolic distribution (GHD). It can handle skewness and heavy tails, and has many well-known distributions as special or limiting cases. The multiple scaled GHD (MSGHD) and the mixture of coalesced GHDs (CGHD) are even more flexible methods that can detect non-elliptical, and even non-convex, clusters. The drawback of high flexibility is a high parametrization — especially so for high-dimensional data because the number of parameters is depends on the number of variables. Therefore, the aforementioned methods are not well suited for high-dimensional data clustering. However, the eigen-decomposition of the component scale matrix can naturally be used for dimension reduction obtaining a transformation of the MSGHD and MCGHD that is better suited for high-dimensional data clustering.

Key words: Mixture models, generalized hyperbolic distribution, cluster analysis, high dimensional data

1 Background: Model-based clustering

Model-based clustering assumes that a population is a convex combination of a finite number of densities. A random vector $\mathbf{X}$ follows a (parametric) finite mixture distribution if, for all $\mathbf{x} \subset \mathbf{X}$, its density can be written

Cristina Tortora
San Jose State University, One Washington square, San Jose CA USA, e-mail: cristina.tortora@sjsu.edu

Paul D. McNicholas
McMaster University, 1280 Main St W, Hamilton, ON Canada e-mail: paul@math.mcmaster.ca

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g),$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, is the gth mixing proportion, $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the gth component density, and $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the vector of parameters, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$. The component densities $f_1(\mathbf{x} \mid \boldsymbol{\theta}_1), \ldots, f_G(\mathbf{x} \mid \boldsymbol{\theta}_G)$ are usually taken to be of the same type. Over the past few years, non-Gaussian model-based clustering techniques have gained popularity. [3] proposed the use of the generalized hyperbolic distribution (GHD), which has the advantage of being extremely flexible because it is characterized by five parameters–the mean, the scale matrix, the skewness, the concentration and the index parameters. Many other distributions, e.g. the Gaussian or the the skew-t distribution, can be obtained as a special or limiting cases. The density of a random variable $\mathbf{X}$ from a generalized hyperbolic distribution is

$$f_{\mathrm{H}}(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{\chi + \boldsymbol{\Sigma}(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})}{\psi + \boldsymbol{\alpha}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}}\right]^{\frac{\lambda - \frac{p}{2}}{2}} \frac{(\psi/\chi)^{\frac{\lambda}{2}} K_{\lambda - \frac{p}{2}}\left(\sqrt{[\psi + \boldsymbol{\alpha}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}][\chi + \boldsymbol{\Sigma}(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})]}\right)}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} K_{\lambda}(\sqrt{\chi\psi}) \exp\{(\boldsymbol{\mu} - \mathbf{x})' \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}\}}, \tag{1}$$

where $\boldsymbol{\Sigma}(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}$, K_{λ} is the modified Bessel function of the third kind with index λ, and $\boldsymbol{\vartheta}$ denotes the parameters. The parameters have the following interpretation: λ is an index parameter, χ and ψ are concentration parameters, $\boldsymbol{\alpha}$ is a skewness parameter, $\boldsymbol{\mu}$ is the mean, and $\boldsymbol{\Sigma}$ is the scale matrix.

Let $Y \sim \mathrm{GIG}(\psi, \chi, \lambda)$, where GIG indicates the generalized inverse Gaussian distribution [1], and the density is given by

$$h(y \mid \boldsymbol{\theta}_g) = \frac{(y/\eta)^{\lambda - 1}}{2\eta K_{\lambda}(\omega)} \exp\left\{-\frac{\omega}{2}\left(\frac{y}{\eta} + \frac{\eta}{y}\right)\right\}. \tag{2}$$

Consider Y and a random variable $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Then, a generalized hyperbolic random variable $\mathbf{X}$, see (1), can be generated via

$$\mathbf{X} = \boldsymbol{\mu} + Y\boldsymbol{\alpha} + \sqrt{Y}\mathbf{V}, \tag{3}$$

and it follows that $\mathbf{X} \mid Y \sim \mathcal{N}(\boldsymbol{\mu} + y\boldsymbol{\alpha}, y\boldsymbol{\Sigma})$.

Note that the parameterization used in (1) requires the constraint $|\boldsymbol{\Sigma}| = 1$ to ensure identifiability, but this constraint is not practical for clustering applications. Therefore, an alternative parameterization, setting $\omega = \sqrt{\psi\chi}$ and $\eta = \sqrt{\chi/\psi}$, is used with $\eta = 1$ (see [3]). Under this parametrization the density of the generalized hyperbolic distribution is

$$f_{\mathrm{H}}(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{\omega + \boldsymbol{\Sigma}(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})}{\omega + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right]^{\frac{\lambda-\frac{p}{2}}{2}} \frac{K_{\lambda-\frac{p}{2}}\left(\sqrt{[\omega + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}][\omega + \boldsymbol{\Sigma}(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})]}\right)}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}K_{\lambda}(\omega)\exp\{-(\boldsymbol{\mu}-\mathbf{x})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\}}. \tag{4}$$

Details of this alternative parameterization, as well as maximum likelihood parameter estimates are given by [3]. Parameter estimation for the mixture of generalized hyperbolic distributions model can be carried out via the expectation-maximization (EM) algorithm [4].

1.1 Multiple scaled generalized hyperbolic distribution

The index and concentration parameters, λ and ω are unidimensional, i.e. they are the same for every dimension. Basing on the idea of [5], [8] proposed the multiple scaled GHD, where $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$ are p-dimensional vectors, i.e., they can vary in each dimension. To introduce the multiple scaled distribution we need to define the normal variance-mean mixture. The distribution of a p-dimensional random variable $\mathbf{X}$ is said to be a normal variance-mean mixture if its density can be written in the form

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \boldsymbol{\theta}) = \int_0^{\infty} \phi_p(\mathbf{x} \mid \boldsymbol{\mu} + w\alpha, f(w)\boldsymbol{\Sigma})\, h(w \mid \boldsymbol{\theta})\, dw, \tag{5}$$

where $\phi_p(\mathbf{x} \mid \boldsymbol{\mu} + w\alpha, w\boldsymbol{\Sigma})$ is the density of a p-dimensional Gaussian distribution with mean $\boldsymbol{\mu} + w\alpha$ and covariance matrix $f(w)\boldsymbol{\Sigma}$, and $h(w \mid \boldsymbol{\theta})$ is the density of a univariate random variable $W > 0$ that has the role of a weight function [2, 6]. This weight function can take on many forms, when the density of W follow a generilized inverse Gaussian distribution, $f_t(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ follows the density of the GHD. [5] show that a multi-dimensional weight variable

$$\boldsymbol{\Sigma}_{\mathbf{W}} = \mathrm{diag}\left(w_1^{-1}, \ldots, w_p^{-1}\right)$$

can be incorporated into (5) via an eigen-decomposition of the symmetric positive-definite matrix $\boldsymbol{\Sigma}$, setting $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'$. Following [7] the formulation of the GHD in (4) can be written as a normal variance-mean mixture where the univariate density is GIG, i.e.,

$$\mathbf{X} = \boldsymbol{\mu} + W\alpha + \sqrt{W}\mathbf{V}, \tag{6}$$

where $\mathbf{V} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and W has density

$$h(w \mid \omega, 1, \lambda) = \frac{w^{\lambda-1}}{2K_{\lambda}(\omega)} \exp\left\{-\frac{\omega}{2}\left(w + \frac{1}{w}\right)\right\}, \tag{7}$$

for $w > 0$, where ω and λ are as previously defined. From (6) and (7), it follows that the generalized hyperbolic density can be written

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \omega, \lambda) = \int_0^\infty \phi_p(\mathbf{x} \mid \boldsymbol{\mu} + w\alpha, w\boldsymbol{\Sigma}) h(w \mid \omega, 1, \lambda) dw. \tag{8}$$

The density of a multiple scaled generalized hyperbolic distribution (MSGHD) is

$$f_{\text{MSGHD}}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \alpha, \omega, \lambda) = \int_0^\infty \cdots \int_0^\infty \phi_p\left(\boldsymbol{\Gamma}'\mathbf{x} - \boldsymbol{\mu} - \Delta_{\mathbf{w}}\alpha \mid \mathbf{0}, \Delta_{\mathbf{w}}\boldsymbol{\Phi}\right) h_{\mathbf{w}}(w_1, \ldots, w_p \mid \omega, \mathbf{1}, \lambda) dw_1 \ldots dw_p, \tag{9}$$

where $\omega = (\omega_1, \ldots, \omega_p)'$, $\lambda = (\lambda_1, \ldots, \lambda_p)'$, $\mathbf{1}$ is a p-vector of 1s, and

$$h_{\mathbf{W}}(w_1, \ldots, w_p \mid \omega, \mathbf{1}, \lambda) = h(w_1 \mid \omega_1, 1, \lambda_1) \times \cdots \times h(w_p \mid \omega_p, 1, \lambda_p).$$

Then, a mixture of MSGHDs (MMSGHDs) has density

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_{\text{MSGHD}}\left(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Gamma}_g, \boldsymbol{\Phi}_g, \alpha_g, \omega_g, \lambda_g\right). \tag{10}$$

Details of maximum likelihood parameter estimates and EM-algorithm are given by [7].

1.2 Mixture of Coalesced Generalized Hyperbolic Distributions

The generalized hyperbolic distribution is not a special or limiting case of the MSGHD under any parameterization with $p > 1$. [7] proposed a coalesced generalized hyperbolic distribution (CGHD) that contains both the generalized hyperbolic distribution and MSGHD as limiting cases. The CGHD arises through the introduction of a random vector

$$\mathbf{R} = U\mathbf{X} + (1-U)\mathbf{S}, \tag{11}$$

where $\mathbf{X} = \boldsymbol{\Gamma}\mathbf{Y}$, $\mathbf{Y} \backsim \text{GHD}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \omega_0, \lambda_0)$, $\mathbf{S} \backsim \text{MSGHD}(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \alpha, \omega, \lambda)$, with $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'$, and U is an indicator variable such that

$$U = \begin{cases} 1 & \text{if } R \text{ follows a generalized hyperbolic distribution, and} \\ 0 & \text{if } R \text{ follows a MSGHD.} \end{cases}$$

It follows that $\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\mu} + W\boldsymbol{\Gamma}\alpha + \sqrt{W}\boldsymbol{\Gamma}\mathbf{V}$, where $\boldsymbol{\Gamma}\mathbf{V} \backsim \text{N}_p\left(\mathbf{0}, \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\right)$, $\mathbf{S} = \boldsymbol{\Gamma}\boldsymbol{\mu} + \boldsymbol{\Gamma}\alpha\Delta_{\mathbf{w}} + \boldsymbol{\Gamma}\mathbf{A}$, where $\boldsymbol{\Gamma}\mathbf{A} \backsim \text{N}_p\left(\mathbf{0}, \boldsymbol{\Gamma}\Delta_{\mathbf{w}}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\right)$, and the density of $\mathbf{R}$ can be written

$$f_{\text{CGHD}}(\mathbf{r} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \alpha, \omega, \lambda, \omega_0, \lambda_0, \varpi) = \varpi f_{\text{GHD}}\left(\mathbf{r} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}', \alpha, \omega_0, \lambda_0\right) + (1-\varpi) f_{\text{MSGHD}}\left(\mathbf{r} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \alpha, \omega, \lambda\right), \tag{12}$$

where $f_{\text{GHD}}(\cdot)$ is the density of a generalized hyperbolic random variable, $f_{\text{MSGHD}}(\cdot)$ is the density of a MSGHD random variable, and $\varpi \in (0,1)$ is a mixing proportion. Note that the random vector $\mathbf{R}$ would be distributed generalized hyperbolic if $\varpi = 1$ and would be distributed MSGHD if $\varpi = 0$.

Parameter estimation can be carried out via a generalized expectation-maximization (GEM) algorithm [4].

2 Dimension reduction

The mixture of GHDs, MSGHDs, and CGHDs are extremely flexible and give good clustering performance; however, the flexibility is obtained increasing the number of parameters. This makes the methods unsuitable for high-dimensional data sets. The problem can be solved considering that the singular value decomposition of the scale matrix $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\Psi\boldsymbol{\Gamma}'$ naturally leads to dimension reduction using $p \times q$ $\boldsymbol{\Gamma}$ and $q \times q$ diagonal $\boldsymbol{\Phi}$ with $q < p$. The random variable $\mathbf{Y}$ is defined as

$$\mathbf{Y} = \boldsymbol{\Gamma}^*\mathbf{X} + \varepsilon, \tag{13}$$

with $\mathbf{X} \sim \text{GHD}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \omega, \lambda)$, and $\varepsilon \sim N(0, \Psi)$ where Ψ is a q dimensional diagonal matrix. Using (11) and (13) It follows that

$$\mathbf{Y} = \boldsymbol{\Gamma}^*\boldsymbol{\mu} + \boldsymbol{\Gamma}^* w\alpha + \boldsymbol{\Gamma}^*\sqrt{w}\mathbf{V} + \varepsilon, \tag{14}$$

and

$$\mathbf{Y} \sim GHD(\boldsymbol{\Gamma}^*\boldsymbol{\mu}, \boldsymbol{\Gamma}^*\alpha, (\boldsymbol{\Gamma}^*\boldsymbol{\Phi}(\boldsymbol{\Gamma}^*)' + \Psi), \lambda, \omega). \tag{15}$$

Similarly if $\mathbf{X} \sim \text{MSGHD}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \omega, \lambda)$,

$$\mathbf{Z} \sim MSGHD(\boldsymbol{\Gamma}^*\boldsymbol{\mu}, \boldsymbol{\Gamma}^*\alpha, \omega, \lambda, \boldsymbol{\Phi} + \Psi). \tag{16}$$

Define $\tilde{\boldsymbol{\Phi}} := \Psi + \boldsymbol{\Phi}$, a $q \times q$ diagonal matrix. The mixture models obtained using the new proposed density function will be defined as low-dimension mixture of GHDs (LMGHDs) and lo- dimension mixture of MSGHDs (LMMSGHDs) respectively. Using the same procedure used in Section 1.2 we can obtain the lowdimension mixture of CGHDs (LMCGHDs). The parameters that maximize the likelihood for each model can be estimated using the EM-algorithm.

References

1. Barndorff-Nielsen, O., Halgreen, C.: Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. Z. Wahrscheinlichkeitstheorie Verw. Gebiete **38**, 309–311 (1977)
2. Barndorff-Nielsen, O., Kent, J., Sørensen, M.: Normal variance-mean mixtures and z distributions. International Statistical Review / Revue Internationale de Statistique **50**(2), 145–159

(1982)
3. Browne, R.P., McNicholas, P.D.: A mixture of generalized hyperbolic distributions. Canadian Journal of Statistics **43**(2), 176–198 (2015)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B **39**(1), 1–38 (1977)
5. Forbes, F., Wraith, D.: A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering. Statistics and Computing **24**(6), 971–984 (2014)
6. Gneiting, T.: Normal scale mixtures and dual probability densities. Journal of Statistical Computation and Simulation **59**(4), 375–384 (1997)
7. Tortora, C., Franczak, B., Browne, R., McNicholas, P.: A mixture of coalesced generalized hyperbolic distributions. Journal of Classification (accepted) (2018)
8. Tortora, C., McNicholas, P.D., Browne, R.P.: A mixture of generalized hyperbolic factor analyzers. Advances in Data Analysis and Classification **10**(4), 423–440 (2016)