O'Hagan, Adrian; **Murphy, Thomas Brendan**; **Gormley, Isobel Claire**; **McNicholas, Paul D.**; **Karlis, Dimitris**

Clustering with the multivariate normal inverse Gaussian distribution. (English)

Zbl 1468.62151
Comput. Stat. Data Anal. 93, 18-30 (2016).

Summary: Many model-based clustering methods are based on a finite Gaussian mixture model. The Gaussian mixture model implies that the data scatter within each group is elliptically shaped. Hence non-elliptical groups are often modeled by more than one component, resulting in model over-fitting. An alternative is to use a mean-variance mixture of multivariate normal distributions with an inverse Gaussian mixing distribution (MNIG) in place of the Gaussian distribution, to yield a more flexible family of distributions. Under this model the component distributions may be skewed and have fatter tails than the Gaussian distribution. The MNIG based approach is extended to include a broad range of eigendecomposed covariance structures. Furthermore, MNIG models where the other distributional parameters are constrained is considered. The Bayesian Information Criterion is used to identify the optimal model and number of mixture components. The method is demonstrated on three sample data sets and a novel variation on the univariate Kolmogorov-Smirnov test is used to assess goodness of fit.

MSC:

62-08 Computational methods for problems pertaining to statistics
62H30 Classification and discrimination; cluster analysis (statistical aspects)

Keywords:

model-based clustering; multivariate normal inverse Gaussian distribution; MCLUST; information metrics; Kolmogorov-Smirnov goodness of fit

Software:

mclust; flowClust; R; mixsmsn

Full Text: DOI Link

References:

[1] Andrews, J. L.; McNicholas, P. D., Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. the teigen family, Stat. Comput., 22, 1021-1029, (2011) · Zbl 1252.62062

[2] Azzalini, A.; Bowman, A. W., A look at some data on the old faithful geyser, Appl. Stat., 39, 357-365, (1990) · Zbl 0707.62186

[3] Banfield, J. D.; Raftery, A. E., Model-based gaussian and non-Gaussian clustering, Biometrics, 49, 803-821, (1993) · Zbl 0794.62034

[4] Baudry, J. P.; Raftery, A.; Celeux, G.; Lo, K.; Gottardo, R., Combining mixture components for clustering, J. Comput. Graph. Statist., 9, 332-353, (2010)

[5] Bensmail, H.; Celeux, G., Regularized Gaussian discriminant analysis through eigenvalue decomposition, J. Amer. Statist. Assoc., 91, 1743-1748, (1996) · Zbl 0885.62068

[6] Cabral, C. R.; Lachos, V. H.; Prates, M., Robust multivariate mixture modelling using scale mixtures of skew-normal distributions, Comput. Statist. Data Anal., 56, 226-246, (2012)

[7] Cabral, C. R.; Lachos, V. H.; Zeller, C. B., Multivariate measurement error models using finite mixtures of skew-student t distributions, J. Multivariate Anal., 124, 179-198, (2014)

[8] Celeux, G.; Govaert, G., Gaussian parsimonious clustering models, Pattern Recognit., 28, 781-793, (1995)

[9] Chen, J.; Hong, H.; Stein, J., Forecasting crashes: trading volume, past returns, and conditional skewness in stock prices, J. Financ. Econ., 61, 345-381, (2001)

[10] Dasgupta, A.; Raftery, A. E., Detecting features in spatial point processes with clutter via model-based clustering, J. Amer. Statist. Assoc., 93, 294-302, (1998) · Zbl 0906.62105

[11] Fisher, R., The use of multiple measurements in taxonomic problems, Ann. Eugenics, 7, 179-188, (1936)

[12] Fraley, C.; Raftery, A. E., How many clusters? which clustering method? answers via model-based cluster analysis, Comput.

J., 41, 578-588, (1998) · Zbl 0920.68038

[13] Fraley, C.; Raftery, A. E., MCLUST: software for model-based clustering, J. Classification, 16, 297-306, (1999) · Zbl 0951.91500

[14] Fraley, C.; Raftery, A. E., Model-based clustering, discriminant analysis, and density estimation, J. Amer. Statist. Assoc., 97, 611-612, (2002) · Zbl 1073.62545

[15] Fruhwirth-Schnatter, S.; Pyne, S., Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions, Biostatistics, 11, 317-336, (2009)

[16] Gottardo, R.; Lo, K., Robust model-based clustering of flow cytometry data: the flowclust package. technical report, (2011), UBC Vancouver, Canada

[17] Hennig, C., Methods for merging Gaussian mixture components, Adv. Data Anal. Classif., 4, 3-34, (2010) · Zbl 1306.62141

[18] Hubert, L.; Arabie, P., Comparing partitions, J. Classification, 2, 193-218, (1985)

[19] Hunter, D.; Wang, S.; Hettmansperger, T., Inference for mixtures of symmetric distributions, Ann. Statist., 35, 224-251, (2007) · Zbl 1114.62035

[20] Karlis, D.; Santourian, A., Model-based clustering with non-elliptically contoured distributions, Stat. Comput., 19, 73-83, (2008)

[21] Kass, R. E.; Raftery, A. E., Bayes factors, J. Amer. Statist. Assoc., 90, 773-795, (1995) · Zbl 0846.62028

[22] King, G., How not to Lie with statistics: avoiding common mistakes in quantitative political science, Amer. J. Polit. Sci., 30, 666-687, (1986)

[23] Lin, T. I.; Ho, H. J.; Chen, C. L., Analysis of multivariate skew normal models with incomplete data, J. Multivariate Anal., 100, 2337-2351, (2009) · Zbl 1175.62054

[24] Lin, T. I.; Lee, J. C.; Hsieh, W. J., Robust mixture modeling using the skew t distribution, Stat. Comput., 17, 81-92, (2007)

[25] Lin, T. I.; Lee, J. C.; Yen, S. Y., Finite mixture modelling using the skew normal distribution, Statist. Sinica, 17, 909-927, (2007) · Zbl 1133.62012

[26] Lin, T. I.; Lin, T., Robust statistical modelling using the multivariate skew t distribution with complete and incomplete data, Stat. Comput., 11, 253-277, (2011) · Zbl 1218.62050

[27] MacLean, C.; Morton, N.; Elston, R.; Yee, S., Skewness in commingling distributions, Biometrics, 32, 695-699, (1976) · Zbl 0334.62012

[28] Massey, F. J., The Kolmogorov-Smirnov test for goodness of fit, J. Amer. Statist. Assoc., 46, 68-78, (1951) · Zbl 0042.14403

[29] McLachlan, G. J.; Peel, D., Finite mixture models, (2000), Wiley Interscience New York · Zbl 0963.62061

[30] McNicholas, S.M., McNicholas, P.D., Browne, R.P., 2013. Mixtures of variance-gamma distributions. ArXiv e-prints arXiv:1309.2695.

[31] McNicholas, P. D.; Murphy, T. B., Model-based clustering of longitudinal data, Canad. J. Statist., 38, 153-168, (2010) · Zbl 1190.62120

[32] Mechel, F., Calculation of the modified Bessel functions of the second kind with complex argument, Math. Comp., 20, 407-412, (1966) · Zbl 0143.17902

[33] Meila, M., Comparing clusterings—an information based distance, J. Multivariate Anal., 98, 873-895, (2007) · Zbl 1298.91124

[34] Mirkin, B. G.; Cherny, L. B., Measurement of the partition between distinct partitions of a finite set of objects, Autom. Remote Control, 31, 786-792, (1970) · Zbl 0221.05029

[35] Prates, M. O.; Cabral, C. R.; Lachos, V. H., Fitting finite mixture of scale mixture of skew-normal distributions, J. Stat. Softw., 54, (2013)

[36] Prates, M.O., Lachos, V.H., Cabral, C.R., 2013b. mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. R package version 0.2-9.

[37] Pyne, S.; Hu, X.; Wang, K.; Rossin, E.; Lin, T. I.; Maier, L. M.; Allan, C. B.; McLachlan, G.; Tamayo, P.; Hafler, D.; De Jager, P. L.; Mesirov, J. P., Automated high-dimensional flow cytometric data analysis, Proc. Natl. Acad. Sci., 106, 8519-8524, (2009)

[38] R Development Core Team, R: A language and environment for statistical computing, (2012), R Foundation for Statistical Computing Vienna, Austria, URL: http://www.R-project.org

[39] Schwarz, G., Estimating the dimension of a model, Ann. Statist., 6, 461-464, (1978) · Zbl 0379.62005

[40] Van Dongen, S., Performance criteria for graph clustering and Markov cluster experiments. technical report, (2000), National Research Institute for Mathematics and Computer Science Amsterdam, Holland

[41] Vrbik, I.; McNicholas, P. D., Parsimonious skew mixture models for model-based clustering and classification, Comput. Statist. Data Anal., 71, 196-210, (2014)

[42] Wang, W. L., Mixtures of common factor analyzers for high-dimensional data with missing information, J. Multivariate Anal., 117, 120-133, (2013) · Zbl 1277.62162

[43] Wasserman, R. E.K. L., A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion, J. Amer. Statist. Assoc., 90, 928-934, (1995) · Zbl 0851.62020

[44] Yu, Y., On normal variance-mean mixtures. technical report, (2011), University of California Irvine, California, USA