Young, Derek S.; **Chen, Xi**; **Hewage, Dilrukshi C.**; **Nilo-Poyanco, Ricardo**

Finite mixture-of-gamma distributions: estimation, inference, and model-based clustering.

(English) Zbl 1474.62043
Adv. Data Anal. Classif., ADAC 13, No. 4, 1053-1082 (2019).

Summary: Finite mixtures of (multivariate) Gaussian distributions have broad utility, including their usage for model-based clustering. There is increasing recognition of mixtures of asymmetric distributions as powerful alternatives to traditional mixtures of Gaussian and mixtures of t distributions. The present work contributes to that assertion by addressing some facets of estimation and inference for mixtures-of-gamma distributions, including in the context of model-based clustering. Maximum likelihood estimation of mixtures of gammas is performed using an expectation-conditional-maximization (ECM) algorithm. The Wilson-Hilferty normal approximation is employed as part of an effective starting value strategy for the ECM algorithm, as well as provides insight into an effective model-based clustering strategy. Inference regarding the appropriateness of a common-shape mixture-of-gammas distribution is motivated by theory from research on infant habituation. We provide extensive simulation results that demonstrate the strong performance of our routines as well as analyze two real data examples: an infant habituation dataset and a whole genome duplication dataset.

MSC:

62F03 Parametric hypothesis testing
62H30 Classification and discrimination; cluster analysis (statistical aspects)
62P15 Applications of statistics to psychology

Keywords:

ECM algorithms; finite mixture models; identifiability; mixturegram; multivariate Gaussian copula; starting values

Software:

GSM; mixtools; lcmix; R; flexmix; mclust

Full Text: DOI

References:

[1] Akaike, H.; Petrov, BN (ed.); Csaki, F. (ed.), Information theory and an extension of the maximum likelihood principle, 267-281 (1973), Budapest

[2] Al-Saleh, JA; Agarwal, SK, Finite mixture of gamma distributions: a conjugate prior, Comput Stat Data Anal, 51, 4369-4378 (2007) · Zbl 1162.62323

[3] Almhana J, Liu Z, Choulakian V, McGorman R (2006) A recursive algorithm for gamma mixture models. In: 2006 IEEE international conference on communications, vol 1, pp 197-202 · Zbl 1431.62338

[4] Atapattu, S.; Tellambura, C.; Jiang, H., A mixture gamma distribution to model the SNR of wireless channels, IEEE Trans Wirel Commun, 10, 4193-4203 (2011)

[5] Baudry, J-P; Celeux, G., EM for mixtures: initialization requires special care, Stat Comput, 25, 713-726 (2015) · Zbl 1331.62301

[6] Benaglia, T.; Chauveau, D.; Hunter, DR; Young, DS, mixtools: an R package for analyzing finite mixture models, J Stat Softw, 32, 1-29 (2009)

[7] Biernaki, C.; Celeux, G.; Govaert, G., Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models, Comput Stat Data Anal, 41, 561-575 (2003) · Zbl 1429.62235

[8] Bochkina, N.; Rousseau, J., Adaptive density estimation based on a mixture of gammas, Electron J Stat, 11, 916-962 (2017) · Zbl 1362.62076

[9] Box, GEP; Cox, DR, An analysis of transformations, J R Stat Soc Ser B, 26, 211-252 (1964) · Zbl 0156.40104

[10] Chen, J.; Kahlili, A., Order selection in finite mixture models with a nonsmooth penalty, J Am Stat Assoc, 104, 187-196 (2009) · Zbl 1388.62040

[11] Chen, H.; Chen, J.; Kalbfleisch, JD, A modified likelihood ratio test for homogeneity in finite mixture models, J R Stat Soc Ser B, 63, 19-29 (2001) · Zbl 0976.62011

[12] Clark, JW; Donoghue, PCJ, Constraining the timing of whole genome duplication in plant evolutionary history, Proc R Soc B Biol Sci, 284, 1-8 (2017)

[13] Clark, JW; Donoghue, PCJ, Whole-genome duplication and plant macroevolution, Trends Plant Sci, 23, 933-945 (2018)

[14] Colombo, J.; Mitchell, DW, Infant visual habituation, Neurobiol Learn Mem, 92, 225-234 (2009)

[15] Colombo, J.; Kapa, L.; Curtindale, L.; Oakes, LM (ed.); Cashon, CH (ed.); Casasola, M. (ed.); Rakison, DH (ed.), Varieties of attention in infancy, 3-26 (2011), New York

[16] Cutler, A.; Cordiero-Braña, OI, Minimum Hellinger distance estimation for finite mixture models, J Am Stat Assoc, 91, 1716-1723 (1996) · Zbl 0881.62035

[17] Dempster, AP; Laird, NM; Rubin, DB, Maximum likelihood from incomplete data via the EM algorithm, J R Stat Soc Ser B, 39, 1-38 (1977) · Zbl 0364.62022

[18] Dvorkin D (2012) lcmix: layered and chained mixture models. R package version 0.3/r5

[19] Evin, G.; Merleau, J.; Perreault, L., Two-component mixtures of normal, gamma, and gumbel distributions for hydrological applications, Water Resour Res, 47, 1-21 (2011)

[20] Feng, ZD; McCulloch, CE, Using bootstrap likelihood ratios in finite mixture models, J R Stat Soc Ser B, 58, 609-617 (1996) · Zbl 0906.62021

[21] Figueiredo, MAT; Jain, AK, Unsupervised learning of finite mixture models, IEEE Trans Pattern Anal Mach Intell, 24, 381-396 (2002)

[22] Fraley, C.; Raftery, AE, Model-based clustering, discriminant analysis, and density estimation, J Am Stat Assoc, 97, 611-631 (2002) · Zbl 1073.62545

[23] Fraley, C.; Raftery, AE, Bayesian regularization for normal mixture estimation and model-based clustering, J Classif, 24, 155-181 (2007) · Zbl 1159.62302

[24] Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Sringer, New York · Zbl 1108.62002

[25] Gárcia-Escudero, LA; Gordaliza, A.; Matrán, C.; Mayo-Iscar, A., A general trimming approach to robust cluster analysis, Ann Stat, 36, 1324-1345 (2008) · Zbl 1360.62328

[26] Gárcia-Escudero, LA; Gordaliza, A.; Matrán, C.; Mayo-Iscar, A., Avoiding spurious local maximizers in mixture modeling, Stat Comput, 25, 619-633 (2015) · Zbl 1331.62100

[27] Gilmore, RO; Thomas, H., Examining individual differences in infants' habituation patterns using objective quantitative techniques, Infant Behav Dev, 25, 399-412 (2002)

[28] Grün, B.; Leisch, F., Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters, J Stat Softw, 28, 1-35 (2008)

[29] Hood, BM; Murray, L.; King, F.; Hooper, R.; Atkinson, J.; Braddick, O., Habituation changes in early infancy: longitudinal measures from birth to 6 months, J Reprod Infant Psychol, 14, 177-185 (1996)

[30] Huang, W-J; Chang, S-H, On some characterizations of the mixture of gamma distributions, J Stat Plan Inference, 137, 2964-2974 (2007) · Zbl 1120.60004

[31] Hubert, L.; Arabie, P., Comparing partitions, J Classif, 2, 193-218 (1985) · Zbl 0587.62128

[32] Ingrassia, S., A likelihood-based constrained algorithm for multivariate normal mixture models, Stat Methods Appl, 13, 151-166 (2004) · Zbl 1205.62066

[33] John, S., On identifying the population of origin of each observation in a mixture of observations from two gamma populations, Technometrics, 12, 565-568 (1970)

[34] Karlis, D.; Xekalaki, E., Minimum Hellinger distance estimation for poisson mixtures, Comput Stat Data Anal, 29, 81-103 (1998) · Zbl 1042.62515

[35] Karlis, D.; Xekalaki, E., Choosing initial values for the EM algorithm for finite mixtures, Comput Stat Data Anal, 41, 577-590 (2003) · Zbl 1429.62082

[36] Kim, D.; Seo, B., Assessment of the number of components in Gaussian mixture models in the presence of multiple local maximizers, J Multivar Anal, 125, 100-120 (2014) · Zbl 1280.62028

[37] Kotz S, Balakrishnan N, Johnson NL (2000) Continuous multivariate distributions, volume 1: models and applications, 2nd edn. Wiley, New York · Zbl 0946.62001

[38] Krishnamoorthy, K.; Mathew, T.; Mukherjee, S., Normal-based methods for a gamma distribution: prediction and tolerance intervals and stress-strength reliability, Technometrics, 50, 69-78 (2008)

[39] Krishnamoorthy, K.; Lee, M.; Xiao, W., Likelihood ratio tests for comparing several gamma distributions, Environmetrics, 26, 571-583 (2015)

[40] Lee, SX; McLachlan, GJ, Model-based clustering and classification with non-normal mixture distributions, Stat Methods Appl, 22, 427-454 (2013) · Zbl 1332.62209

[41] Li, Z.; Baniaga, AE; Sessa, EB; Scascitelli, M.; Graham, SW; Rieseberg, LH; Barker, MS, Early genome duplications in conifers and other seed plants, Sci Adv, 1, 1-8 (2015)

[42] Li, H-C; Kyrlov, VA; Fan, P-Z; Zerubia, J.; Emery, WJ, Unsupervised learning of generalized gamma mixture model with application in statistical modeling of high-resolution SAR images, IEEE Trans Geosci Remote Sens, 54, 2153-2170 (2016)

[43] Lindsay, BG, Efficiency versus robustness: the case for minimum Hellinger distance estimation and related methods, Ann Stat, 22, 1081-1114 (1994) · Zbl 0807.62030

[44] Lindsay BG (1995) Mixture models: theory, geometry and applications, volume 5 of NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics and the American Statistical Association

[45] Manly, BFJ, Exponential data transformations, J R Stat Soc Ser D (Stat), 25, 37-42 (1976)

[46] Mathai, AM; Moschopoulos, PG, A form of multivariate gamma distribution, Ann Inst Stat Math, 44, 97-106 (1992) · Zbl 0760.62049

[47] Mayrose, I.; Friedman, N.; Pupko, T., A gamma mixture model better accounts for among site rate heterogeneity, Bioinformatics, 21, 151-158 (2005)

[48] McLachlan, GJ, On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture, Appl Stat, 36, 318-324 (1987)

[49] McLachlan, GJ, On the choice of starting values for the EM algorithm in fitting finite mixture models, J R Stat Soc Ser D, 37, 1988 (1988)

[50] McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York · Zbl 0963.62061

[51] McNicholas PD (2016) Mixture model-based classification. CRC Press, Boca Raton · Zbl 1454.62005

[52] Meng, X-L; Rubin, DB, Maximum likelihood estimation via the ECM algorithm: a general framework, Biometrika, 80, 267-278 (1993) · Zbl 0778.62022

[53] Nei, M.; Gojobori, T., Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, Mol Biol Evol, 3, 418-426 (1986)

[54] Nemec, J.; Linnell-Nemec, AF, Mixture models for studying stellar populations I. Univariate mixture models, parameter estimation, and the number of discrete population components, Publ Astron Soc Pac, 103, 95-121 (1991)

[55] Nielsen F (2012) K-MLE: a fast algorithm for learning statistical mixture models. In: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 869-872

[56] Nwe TL, Nguyen TH, Ma B (2014) On the use of Bhattacharyya based GMM distance and neural net features for identification of cognitive load levels. In: INTERSPEECH 2014, 15th annual conference of the international speech communication association, pp 736-740

[57] Pagel, M.; Meade, A., A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data, Syst Biol, 53, 571-581 (2004)

[58] Panchy, N.; Lehti-Shiu, M.; Shiu, S-H, Evolution of gene duplication in plants, Plant Physiol, 171, 2294-2316 (2016)

[59] R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

[60] Ruppert D (2001) Multivariate transformations. In: El-Shaarawi AH, Piegorsch WW (eds) Encyclopedia of environmetrics. Wiley, New York

[61] Schwander, O.; Nielsen, F.; Handcock, E. (ed.); Pelilo, M. (ed.), Fast learning of gamma mixture models with \(k\)-mle, No. 7953, 235-249 (2013), Berlin

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

[63] Scrucca, L.; Fop, M.; Murphy, TB; Raftery, AE, Mclust5: clustering, classification and density estimation using Gaussian finite mixture models, R J, 8, 289-317 (2016)

[64] Sfikas, G.; Constantinopoulos, C.; Likas, A.; Galatsanos, NP; Duch, W. (ed.); Kacprzyk, J. (ed.); Oja, E. (ed.); Zadrożny, S. (ed.), An analytic distance metric for Gaussian mixture models with application in image retrieval, No. 3697, 835-840 (2005), Berlin

[65] Slater, A., Can measures of infant habituation predict later intellectual ability?, Arch Dis Child, 77, 474-476 (1997)

[66] Song, PX-K, Multivariate dispersion models generated from Gaussian copulas, Scand J Stat, 27, 305-320 (2000) · Zbl 0955.62054

[67] Thomas, H.; Faßbender, I., Modeling infant \(i\)'s look on trial \(t\): race-face preference depends on \(i\)'s looking style, Front Psychol, 8, 1-11 (2017)

[68] Thomas, H.; Hettmansperger, TP, Modelling change in cognitive understanding with finite mixtures, J R Stat Soc Ser C, 50, 435-448 (2001) · Zbl 1112.62331

[69] Thomas, H.; Lohaus, A.; Domsch, H.; Hunter, DR (ed.); Richards, DSP (ed.); Rosenberger, JL (ed.), Extensions of reliability theory, 309-316 (2011), Singapore · Zbl 1414.62492

[70] Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York · Zbl 0646.62013

[71] Todd, RT; Forche, A.; Selmecki, A., Ploidy variation in fungi: polyploidy, aneuploidy, and genome evolution, Microbiol Spect, 5, 1-31 (2017)

[72] Vaidyanathan, VS; Vani Lakshmi, R., Estimation of parameters in a finite mixture of multivariate gamma distributions using Gaussian approximation, Sri Lankan J Appl Stat, 17, 187-200 (2016)

[73] Peer, Y.; Fawcett, JA; Proost, S.; Sterck, L.; Vandepoele, K., The flowering world: a tale of duplications, Trends Plant Sci, 14, 680-688 (2009)

[74] Peer, Y.; Mizrachi, E.; Marchal, K., The evolutionary significance of polyploidy, Nat Rev Genet, 18, 411-424 (2017)

[75] Vani Lakshmi, R.; Vaidyanathan, VS, Parameter estimation in gamma mixture model using normal-based approximation, J Stat Theory Appl, 15, 25-35 (2016)

[76] Vanneste, K.; Peer, Y.; Maere, S., Inference of genome duplications from age distributions revisited, Mol Biol Evol, 30, 177-190 (2013)

[77] Vardi, Y.; Shepp, LA; Kaufman, L., A statistical model for positron emission tomography, J Am Stat Assoc, 80, 8-20 (1985) · Zbl 0561.62094

[78] Venturini, S.; Dominici, F.; Parmigiani, G., Gamma shape mixtures for heavy-tailed distributions, Ann Appl Stat, 2, 756-776 (2008) · Zbl 1400.62292

[79] Walker, JF; Yang, Y.; Feng, T.; Timoneda, A.; Mikenas, J.; Hutchison, V.; Edwards, C.; Wang, N.; Ahluwalia, S.; Olivieri, J.; Walker-Hale, N.; Majure, LC; Puente, R.; Kadereit, G.; Lauterbach, M.; Eggli, U.; Flores-Olvera, H.; Ochoterena, H.; Brockington, SF; Moore, MJ; Smith, SA, From cacti to carnivores: improved phylotranscriptomic sampling and hierarchical homology inference provide further insight into the evolution of caryophyllales, Am J Bot, 105, 446-462 (2018)

[80] Wilson, EB; Hilferty, MM, The distribution of chi-square, Proc Natl Acad Sci, 17, 684-688 (1931) · Zbl 0004.36005

[81] Wiper, M.; Insua, DR; Ruggeri, F., Mixtures of gamma distributions with applications, J Comput Graph Stat, 10, 440-454 (2001)

[82] Woodward, WA; Parr, WC; Schucany, WR; Lindsey, H., A comparison of minimum distance and maximum likelihood estimation of a mixture proportion, J Am Stat Assoc, 79, 590-598 (1984) · Zbl 0547.62017

[83] Xu, L.; Jordan, M., On convergence properties of the EM algorithm for Gaussian mixtures, Neural Comput, 8, 129-151 (1996)

[84] Yang, Y.; Moore, MJ; Brockington, SF; Mikenas, J.; Olivieri, J.; Walker, JF; Smith, SA, Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in caryophyllales, including two allopolyploidy events, New Phytol, 217, 855-870 (2018)

[85] Young, DS; Hunter, DR, Random effects regression mixtures for analyzing infant habituation, J Appl Stat, 42, 1421-1441 (2015)

[86] Young, DS; Ke, C.; Zeng, X., The mixturegram: a visualization tool for assessing the number of components in finite mixture models, J Comput Graph Stat, 27, 564-575 (2018)