Wallace, Meredith L.; **Buysse, Daniel J.**; **Germain, Anne**; **Hall, Martica H.**; **Iyengar, Satish**
Variable selection for skewed model-based clustering: application to the identification of novel sleep phenotypes. (English) Zbl 1398.62347
J. Am. Stat. Assoc. 113, No. 521, 95-110 (2018).

Summary: In sleep research, applying finite mixture models to sleep characteristics captured through multiple data types, including self-reported sleep diary, a wrist monitor capturing movement (actigraphy), and brain waves (polysomnography), may suggest new phenotypes that reflect underlying disease mechanisms. However, a direct mixture model application is challenging because there are many sleep variables from which to choose, and sleep variables are often highly skewed even in homogenous samples. Moreover, previous sleep research findings indicate that some of the most clinically interesting solutions will be those that incorporate all three data types. Thus, we present two novel skewed variable selection algorithms based on the multivariate skew normal (MSN) distribution: one that selects the best set of variables ignoring data type and another that embraces the exploratory nature of clustering and suggests multiple statistically plausible sets of variables that each incorporate all data types. Through a simulation study, we empirically compare our approach with other asymmetric and normal dimension reduction strategies for clustering. Finally, we demonstrate our methods using a sample of older adults with and without insomnia. The proposed MSN-based variable selection algorithm appears to be suitable for both MSN and multivariate normal cluster distributions, especially with moderate to large-sample sizes.

MSC:

62P10 Applications of statistics to biology and medical sciences; meta analysis
62H30 Classification and discrimination; cluster analysis (statistical aspects)
62F07 Statistical ranking and selection procedures

Cited in **4** Documents

Keywords:

insomnia; mixture model; research domain criteria; skewed data

Software:

EMMIXcskew; mixture; mclust; Emmixuskew; R; sparcl; PGMM; clustvarsel

Full Text: DOI Link

References:

[1] \textit{Diagnostic and Statistical Manual of Mental Disorders}, (2013)

[2] Andrews, J. L.; McNicholas, P. D., Variable selection for classification and clustering, \textit{Journal of Classification}, 31, 136-153, (2013) · Zbl 1360.62310

[3] \textit{vscc: Variable Selection for Clustering and Classification}, (2013)

[4] Azzalini, A., \textit{The R Package : The Skew-Normal and Skew-\textit{t} Distributions (version 1.1-2)}, (2014), Università di Padova, Italia

[5] Azzalini, A.; Genton, M. G., Robust likelihood methods based on the skew-\textit{t} and related distributions, \textit{International Statistical Review}, 76, 106-129, (2008) · Zbl 1206.62102

[6] Azzalini, A.; Valle, A. D., Multivariate skew-normal distribution, \textit{Biometrika}, 83, 715-726, (1996) · Zbl 0885.62062

[7] Bafadhel, M.; McKenna, S.; Terry, S.; Mistry, V.; Reid, C.; Haldar, P.; McCormick, M.; Haldar, K.; Kebadze, T.; Duvoix, A.; Lindblad, K.; Patel, H.; Rugman, P.; Dodson, P.; Jenkins, M.; Saunders, M.; Newbold, P.; Green, R. H.; Venge, P.; Lomas, D. A.; Barer, M. R.; Johnston, S. L.; Pavord, I. D.; Brightling, C. E., Acute exacerbations of chronic obstructive pulmonary disease: identification of biologic clusters and their biomarkers, \textit{American Journal of Respiratory and Critical Care Medicine}, 184, 662-671, (2011)

[8] Baillet, M.; Cosin, C.; Schweitzer, P.; Pérés, K.; Catheline, G.; Swendsen, J.; Mayo, W., Mood influences the concordance of subjective and objective measures of sleep duration in older adults, \textit{Frontiers in Aging Neuroscience}, 8, 181, (2016)

[9] Baudry, J.; Raftery, A. E.; Celeux, G.; Lo, K.; Gottardo, R., Combining mixture components for clustering, \textit{Journal of Computational and Graphical Statistics}, 19, 332-353, (2010)

[10] Borodulin, K.; Evenson, K. R.; Monda, K.; Wen, F.; Herring, A. H.; Dole, N., Physical activity and sleep among pregnant

women, \textit{Paediatric and Perinatal Epidemiology}, 24, 45-52, (2009)

[11] Bouveyron, C.; Brunet, C., Simultaneous model-based clustering and visualization in the Fisher discriminative subspace, \textit{Statistics and Computing}, 22, 301-324, (2012) · Zbl 1322.62162

[12] Browne, R. P.; McNicholas, P. D., A mixture of generalized hyperbolic distributions, \textit{Canadian Journal of Statistics}, 43, 176-198, (2015) · Zbl 1320.62144

[13] Buysse, D. J.; Germain, A.; Moul, D. E.; Franzen, P. L.; Brar, L. K.; Fletcher, M. E.; Begley, A.; Houck, P. R.; Mazumdar, S.; Reynolds, C. F.; Monk, T. H., Efficacy of brief behavioral treatment for chronic insomnia in older adults, \textit{Archives of Internal Medicine}, 171, 887-895, (2011)

[14] Cabral, C.; Lachos, V. H.; Prates, M. O., Multivariate mixture modeling using skew-normal independent distributions, \textit{Computational Statistics and Data Analysis}, 56, 126-142, (2012) · Zbl 1239.62058

[15] Casey, B. J.; Craddock, N.; Cuthbert, B. N.; Hyman, S. E.; Lee, F. S.; Ressler, K. J., DSM-5 and rdoc: progress in psychiatry research?, \textit{Nature Reviews Neuroscience}, 14, 810-814, (2013)

[16] Ciu, Y.; Fern, X. Z.; Dy, J. G., Non-redundant multi-view clustering via orthogonalization, \textit{Seventh IEEE International Conference on Data Mining}, 133-142, (2007)

[17] Cuthbert, B. N.; Insel, T. R., Toward the future of psychiatric diagnosis: the seven pillars of rdoc, \textit{BMC Medicine}, 11, 126, (2013)

[18] Dew, M. A.; Hoch, C. C.; Buysee, D. J.; Monk, T. M.; Begley, A. E.; Houck, P. R.; Hall, M. H.; Kupfer, D. J.; Reynolds, C. F., Healthy older adults' sleep predicts all-cause mortality at 4–19 years of follow-up, \textit{Psychosomatic Medicine}, 65, 63-73, (2003)

[19] Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D., \textit{Proceedings of the National Academy of Sciences of the United States of America}, 95, Cluster analysis and display of genome-wide expression patterns, 14836-14868, (1998)

[20] Fakhry, C.; Markis, M. A.; Gilman, R. H.; Cabrerra, L.; Yori, P.; Kosek, M.; Gravitt, P. E., Comparison of the immune microenvironment of the oral cavity and cervix in healthy women, \textit{Cytokine}, 64, 597-604, (2013)

[21] Fraley, C.; Raftery, A. E., How many clusters? which clustering method? answers via model-based cluster analysis, \textit{Computer Journal}, 41, 578-588, (1998) · Zbl 0920.68038

[22] Fraley, C.; Raftery, A. E.; Murphy, T. B.; Scrugga, L., Mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation, (2012)

[23] Franczac, B. C.; Browne, R. P.; McNicholas, P. D., Mixtures of shifted asymmetric Laplace distributions, \textit{IEEE Transactions on Pattern Analysis and Machine Intelligence}, 36, 1149-1157, (2014)

[24] Harvey, A.; Tang, N., (mis)perception of sleep in insomnia: A puzzle and a resolution, \textit{Psychological Bulletin}, 138, 77-101, (2012)

[25] Hauri, P. J.; Sateia, M. J. (eds.), \textit{International Classification of Sleep Disorders: Diagnostic and Coding Manual}, (2005)

[26] Hennig, C., Methods for merging Gaussian mixture components, \textit{Advances in Data Analysis and Classification}, 4, 3-34, (2010) · Zbl 1306.62141

[27] Hlebowics, J.; Persson, M.; Gullberg, B.; Sonestedt, E.; Wallstrom, P.; Drake, I.; Nilsson, J.; Hedblad, B.; Wirfalt, E., Food patterns, inflammation markers and incidence of cardiovascular disease: the malmo diet and cancer study, \textit{Journal of Internal Medicine}, 270, 365-376, (2011)

[28] Hubert, L.; Arabie, P., Comparing partitions, \textit{Journal of Classification}, 2, 193-218, (1984)

[29] Insel, T.; Cuthbert, B.; Garvey, B.; Heinssen, R.; Pine, D. S.; Quinn, K.; Sanislow, C.; Wang, P., Research domain criteria (rdoc): toward a new classification framework for research on mental disorders, \textit{American Journal of Psychiatry}, 167, 748-51, (2010)

[30] Karlis, D.; Meligkotsidou, L., Finite mixtures of multivariate Poisson distributions with application, \textit{Journal of Statistical Planning and Inference}, 137, 1942-1960, (2007) · Zbl 1116.60006

[31] Karlis, D.; Santourian, A., Model-based clustering with non-elliptically contoured distributions, \textit{Statistics and Computing}, 19, 73-83, (2009)

[32] Kass, R. E.; Raftery, A. E., Bayes factors, \textit{Journal of the American Statistical Association}, 90, 773-795, (1995) · Zbl 0846.62028

[33] Kay, D. B.; Buysse, D. J.; Germain, A.; Hall, M. H.; Monk, T. H., Subjective-objective sleep discrepancy among older adults: associations with insomnia diagnosis and insomnia treatment, \textit{Journal of Sleep Research}, 24, 32-39, (2015)

[34] Lachos, V. H.; Ghosh, P.; Arellano-Valle, R. B., Likelihood-based inference for skew-normal independent linear mixed models, \textit{Statistica Sinica}, 20, 303-322, (2010) · Zbl 1186.62071

[35] Lee, S. X.; McLachlan, G. J., Emmixuskew: an R package for Fitting mixtures of multivariate skew \textit{t}-distributions via the EM algorithm, \textit{Journal of Statistical Software}, 55, 1-22, (2013)

[36] "model-based clustering and classification with non-normal mixture distributions" (with discussion), \textit{Statistical Methods and Applications}, 22, 427-479, (2013) · Zbl 1332.62210

[37] On mixtures of skew normal and skew t-distributions, \textit{Advances in Data Analysis and Classification}, 7, 241-266, (2013) · Zbl 1273.62115

[38] Finite mixtures of multivariate skew \textit{t}-distributions: some recent and new results, \textit{Statistics and Computing}, 24, 181-202, (2014) · Zbl 1325.62107

[39] Emmixcskew: an R package for the Fitting of a mixture of canonical fundamental skew \textit{t}-distributions, (2016)

[40] Finite mixtures of canonical fundamental skew \textit{t}-distributions: the unification of the restricted and unrestricted skew \textit{t}-mixture models, \textit{Statistics and Computing}, 26, 573-589, (2016) · Zbl 1420.60020

[41] Levenson, J. C.; Kay, D. B.; Buysse, D. J., The pathophysiology of insomnia, \textit{Chest}, 147, 1179-1192, (2015)

[42] Lin, T.; McLachlan, G. J.; Lee, S. X., A robust factor analysis model using the restricted skew-\textit{t} distribution, \textit{TEST}, 24, 510-531, (2015) · Zbl 1327.62344

[43] Extending mixtures of factor models using the restricted multivariate skew-normal distribution, \textit{Journal of Multivariate Analysis}, 143, 438-318, (2016)

[44] Lin, T. I., Maximum likelihood estimation for multivariate skew normal mixture models, \textit{Journal of Multivariate Analysis}, 100, 257-265, (2009) · Zbl 1152.62034

[45] Robust mixture modeling using multivariate skew t-distributions, \textit{Statistics and Computing}, 20, 343-356, (2010)

[46] Lo, K.; Gottardo, R., Flexible mixture modeling via the multivariate \textit{t} distribution with the box–Cox transformation: an alternative to the skew-\textit{t} distribution, \textit{Statistics and Computing}, 22, 32-52, (2012) · Zbl 1322.62173

[47] Lund, H. G.; Rybarczyk, B. D.; Perrin, P. B.; Leszczysyzn, D.; Stepanski, E., The discrepancy between subjective and objective measures of sleep in older adults receiving CBT for comorbid insomnia, \textit{Journal of Clinical Psychology}, 69, 1108-1120, (2013)

[48] Maugis, C.; Celeux, G.; Martin-Magneitte, M., Variable selection for clustering with Gaussian mixture models, \textit{Biometrics}, 65, 701-709, (2009) · Zbl 1172.62021

[49] McLachlan, G. J.; Bean, R. W.; Peel, D., A mixture model-based approach to the clustering of micro-array expression data, \textit{Bioinformatics}, 18, 413-422, (2002)

[50] McNicholas, P. D.; ElSherbiny, A.; McDaid, A. F.; Murphy, T. B., \textit{pgmm: Parsimonious Gaussian Mixture Models}, (2015)

[51] McNicholas, P. D.; Murphy, T. B., Parsimonious Gaussian mixture models, \textit{Statistics and Computing}, 18, 285-296, (2008)

[52] Model-based clustering of microarray expression data via latent Gaussian mixture models, \textit{Bioinformatics}, 26, 2705-2712, (2010) · Zbl 1203.82150

[53] Monk, T. H.; Reynods, C. F.; Kupfer, D. J.; Buysse, D. J.; Coble, P. A.; Hayes, A. J.; Machen, M. A.; Petrie, S. R.; Ritenour, A. M., The Pittsburgh sleep diary, \textit{Journal of Sleep Research}, 3, 111-120, (1994)

[54] Murray, P. M.; Browne, R. B.; McNicholas, P. D., Mixtures of skew-t factor analyzers, \textit{Computational Statistics and Data Analysis}, 77, 326-335, (2014) · Zbl 06984029

[55] Murray, P. M.; McNicholas, P. D.; Browne, R. B., A mixture of common skew-\textit{t} factor analyzers, \textit{Stat}, 3, 68-82, (2014)

[56] Pyne, S.; Hu, S.; Wang, K.; Rossin, E.; Lin, T.; Maier, L. M.; Baecher-Allan, C.; McLachlan, G. J.; Tamoyo, P.; Hafler, D. A.; De Jager, P. L.; Mesirov, J. P., Automated high-dimensional flow cytometric data analysis, \textit{Proceedings of the National Academy of Sciences in the United States of America}, 106, 8519-8524, (2009)

[57] \textit{R: A Language and Environment for Statistical Computing}, (2015), R Foundation for Statistical Computing, Vienna, Austria

[58] Raftery, A. E.; Dean, N., Variable selection for model-based clustering, \textit{Journal of the American Statistical Association}, 101, 168-178, (2006) · Zbl 1118.62339

[59] Reinke, S.; Broadhurst, D.; Sykes, B.; Baker, G. B.; Catz, I.; Warren, K. G.; Power, C., Metabolomic profiling in multiple sclerosis: insights into biomarkers and pathogenesis, \textit{Multiple Sclerosis Journal}, 20, 1396-1400, (2014)

[60] Sahu, S. K.; Dey, D. K.; Branco, M. D., A new class of multivariate skew distributions with applications to Bayesian regression models, \textit{Canadian Journal of Statistics}, 31, 129-150, (2003) · Zbl 1039.62047

[61] Schork, N. J.; Weder, A. B.; Schork, A., On the asymmetry of biological frequency distributions, \textit{Genetic Epidemioligy}, 7, 427-446, (1990)

[62] Scrucca, L.; Raftery, A. E., Clustvarsel: A package implementing variable selection for model-based clustering in R, (2014)

[63] Steinley, D., Properties of the hubert–arabie adjusted rand index, \textit{Psychological Methods}, 9, 386-396, (2004)

[64] Tarokh, L.; Carskadon, M. A.; Achermann, P., Trait-like characteristics of the sleep EEG across adolescent development, \textit{The Journal of Neuroscience}, 31, 6371-6378, (2011)

[65] Tortora, C.; Franczak, B. C.; Browne, R. P.; ElSherbiny, A.; McNicholas, P. D., A mixture of coalesced generalized hyperbolic distributions, (2014)

[66] Tortora, C.; McNicholas, P.; Browne, R., A mixture of generalized hyperbolic factor analyzers, \textit{Advances in Data Analysis and Classification}, 10, 423-440, (2016) · Zbl 1414.62278

[67] Troxel, W. M.; Buysse, D. J.; Matthews, K. A.; Kip, K. E.; Strollo, P. J.; Hall, M.; Drumheller, O.; Reis, S. E., Sleep symptoms predict the development of the metabolic syndrome, \textit{Sleep}, 33, 1633-40, (2010)

[68] Troxel, W. M.; Kupfer, D. J.; Reynolds, C. F.; Frank, E.; Thase, M. E.; Miewald, J. M.; Buysse, D. J., Insomnia and objectively measured sleep disturbances predict treatment outcome in depressed patients treated with psychotherapy or psychotherapy–pharmacotherapy combinations, \textit{Journal of Clinical Psychiatry}, 73, 478-85, (2012)

[69] Vgontzas, A. N.; Fernandez-Mendoza, J. F.; Liao, D.; Bixler, E. O., Insomnia with objective short sleep duration: the most biologically severe phenotype?, \textit{Sleep Medicine Reviews}, 7, 241-254, (2013)

[70] Vrbik, I.; McNicholas, P. D., Analytic calculations for the EM algorithm for multivariate skew t-mixture models, \textit{Statistics and Probability Letters}, 82, 1169-1174, (2012) · Zbl 1244.65012

[71] Wang, K.; Ng, A.; McLachlan, G., \textit{EMMIXskew: The EM Algorithm and Skew Mixture Distribution}, (2013)

[72] Witten, D. M.; Tibshirani, R., A framework for feature selection in clustering, \textit{Journal of the American Statistical Association}, 105, 713-726, (2010) · Zbl 1392.62194

[73] \textit{sparcl: Perform Sparse Hierarchical Clustering and Sparse k-Means Clustering}, (2013)

[74] Wraith, D.; Forbes, F., Location and scale mixtures of gaussians with flexible tail behavior: properties, inference, and application to multivariate clustering, \textit{Computational Statistics and Data Analysis}, 90, 61-73, (2015) · Zbl 1468.62210

[75] Yeung, K. Y.; Fraley, C.; Murua, A.; Raftery, A. E.; Ruzzo, W. L., Model based clustering and data transformations for gene expression data, \textit{Bioinformatics}, 17, 977-987, (2001)