Watanabe, Chihiro; **Suzuki, Taiji**

Selective inference for latent block models. (English) Zbl 1471.62265

Electron. J. Stat. 15, No. 1, 3137-3183 (2021).

Summary: Model selection in latent block models has been a challenging but important task in the field of statistics. Specifically, a major challenge is encountered when constructing a test on a block structure obtained by applying a specific clustering algorithm to a finite size matrix. In this case, it becomes crucial to consider the selective bias in the block structure, that is, the block structure is selected from all the possible cluster memberships based on some criterion by the clustering algorithm. To cope with this problem, this study provides a selective inference method for latent block models. Specifically, we construct a statistical test on a set of row and column cluster memberships of a latent block model, which is given by a squared residue minimization algorithm. The proposed test, by its nature, includes and thus can also be used as the test on the set of row and column cluster numbers. We also propose an approximated version of the test based on simulated annealing to avoid combinatorial explosion in searching the optimal block structure. The results show that the proposed exact and approximated tests work effectively, compared to the naive test that did not take the selective bias into account.

MSC:

62F03 Parametric hypothesis testing
62H30 Classification and discrimination; cluster analysis (statistical aspects)

Keywords:

latent block model; selective inferenc; relational data analysis

Software:

Bsig; MovieLens

Full Text: DOI arXiv

References:

[1] Basu, D. (1955). On statistics independent of a complete sufficient statistic. \textit{Sankhyā: The Indian Journal of Statistics} 15 377-380. · Zbl 0068.13401

[2] Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. \textit{The Annals of Statistics} 41 802-837. · Zbl 1267.62080

[3] Bickel, P. J. and Sarkar, P. (2016). Hypothesis testing for automated community detection in networks. \textit{Journal of the Royal Statistical Society: Series B (Statistical Methodology)} 78 253-273. · Zbl 1411.62162

[4] Chi, E. C., Allen, G. I. and Baraniuk, R. G. (2017). Convex biclustering. \textit{Biometrics} 73 10-19. · Zbl 1366.62208

[5] Cho, H., Dhillon, I. S., Guan, Y. and Sra, S. (2004). Minimum sum-squared residue co-clustering of gene expression data. In \textit{Proceedings of the 2004 SIAM International Conference on Data Mining} 114-125.

[6] Conover, W. J. (1999). \textit{Practical nonparametric statistics}. John Wiley \& Sons, New York.

[7] Fithian, W., Sun, D. and Taylor, J. (2014). Optimal inference after model selection. arXiv:1410.2597.

[8] Gangrade, A., Venkatesh, P., Nazer, B. and Saligrama, V. (2019). Efficient near-optimal testing of community changes in balanced stochastic block models. In \textit{Advances in Neural Information Processing Systems 32} 10364-10375.

[9] Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. \textit{Pattern Recognition} 36 463-473. · Zbl 1452.62444

[10] Hajek, B. (1988). Cooling schedules for optimal annealing. \textit{Mathematics of Operations Research} 13 311-329. · Zbl 0652.65050

[11] Harper, F. M. and Konstan, J. A. (2015). The MovieLens datasets: history and context. \textit{ACM Transactions on Interactive Intelligent Systems} 5 1-19.

[12] Hartigan, J. A. (1972). Direct clustering of a data matrix. \textit{Journal of the American Statistical Association} 67 123-129.

[13] Henriques, R. and Madeira, S. C. (2018). BSig: evaluating the statistical significance of biclustering solutions. \textit{Data Mining and Knowledge Discovery} 32 124-161. · Zbl 1416.62340

[14] Hu, J., Zhang, J., Qin, H., Yan, T. and Zhu, J. (2020). Using maximum entry-wise deviation to test the goodness-of-fit for stochastic block models. \textit{Journal of the American Statistical Association} 0 1-30.

[15] Inoue, S., Umezu, Y., Tsubota, S. and Takeuchi, I. (2017). Post clustering inference for heterogeneous data. In \textit{Information-Based Induction Science Workshop} 69-76.

[16] Karwa, V., Pati, D., Petrović, S., Solus, L., Alexeev, N., Raič, M., Wilburne, D., Williams, R. and Yan, B. (2016). Exact tests for stochastic block models. arXiv:1612.06040.

[17] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. \textit{Science} 220 671-680. · Zbl 1225.90162

[18] Lee, J. D., Sun, Y. and Taylor, J. E. (2015). Evaluating the statistical significance of biclusters. In \textit{Advances in Neural Information Processing Systems 28} 1324-1332.

[19] Lee, J. D. and Taylor, J. E. (2014). Exact post model selection inference for marginal screening. In \textit{Advances in Neural Information Processing Systems 27} 136-144.

[20] Lee, M., Shen, H., Huang, J. Z. and Marron, J. S. (2010). Biclustering via sparse singular value decomposition. \textit{Biometrics} 66 1087-1095. · Zbl 1233.62182 · doi:10.1111/j.1541-0420.2010.01392.x

[21] Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. \textit{The Annals of Statistics} 44 907-927. · Zbl 1341.62061

[22] Lei, J. (2016). A goodness-of-fit test for stochastic block models. \textit{The Annals of Statistics} 44 401-424. · Zbl 1331.62283

[23] Loftus, J. R. and Taylor, J. E. (2015). Selective inference in regression models with groups of variables. arXiv:1511.01478.

[24] Lomet, A., Govaert, G. and Grandvalet, Y. (2012). Model selection in block clustering by the integrated classification likelihood. In \textit{Proceedings of the 20th International Conference on Computational Statistics} 519-530. · Zbl 1416.62349

[25] Nadif, M. and Govaert, G. (2010). Model-based co-clustering for continuous data. In \textit{Proceedings of the 9th International Conference on Machine Learning and Applications} 175-180. · Zbl 1187.62117

[26] Perrone, V., Jenkins, P. A., Spanò, D. and Teh, Y. W. (2017). Poisson random fields for dynamic feature models. \textit{Journal of Machine Learning Research} 18 1-45. · Zbl 1442.62070

[27] Saber, H. B., Elloumi, M. and Nadif, M. (2011). Block mixture model for the biclustering of microarray data. In \textit{Proceedings of the 22nd International Workshop on Database and Expert Systems Applications} 423-427.

[28] Shan, H. and Banerjee, A. (2008). Bayesian co-clustering. In \textit{Proceedings of the 8th IEEE International Conference on Data Mining} 530-539.

[29] Shao, J. (2003). \textit{Mathematical Statistics}. Springer-Verlag New York.

[30] Tan, K. M. and Witten, D. M. (2014). Sparse biclustering of transposable data. \textit{Journal of computational and graphical statistics} 23 985-1008.

[31] Terada, Y. and Shimodaira, H. (2017). Selective inference for the problem of regions via multiscale bootstrap. arXiv:1711.00949.

[32] Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. \textit{The Annals of Statistics} 46 679-710. · Zbl 1392.62144

[33] Tibshirani, R. J., Rinaldo, A., Tibshirani, R. and Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. \textit{The Annals of Statistics} 46 1255-1287. · Zbl 1392.62210

[34] Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. \textit{Journal of Optimization Theory and Applications} 45 41-51. · Zbl 0534.90091

[35] Watanabe, C. and Suzuki, T. (2021). Goodness-of-fit test for latent block models. \textit{Computational Statistics \& Data Analysis} 154 107090. · Zbl 07345038

[36] Wyse, J. and Friel, N. (2012). Block clustering with collapsed latent block models. \textit{Statistics and Computing} 22 415-428. · Zbl 1322.62046

[37] Yuan, M., Feng, Y. and Shang, Z. (2018). A likelihood-ratio type test for stochastic block models with bounded degrees. arXiv:1807.04426.