НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

НАУЧНЫЙ ДОКЛАД

по результатам подготовленной
научно-квалификационной работы (диссертации)
на тему «Статистическая неопределённость алгоритмов идентификации графических моделей»

ФИО Гречихин Иван Сергеевич

Направление подготовки 09.06.01 Информатика и вычислительная техника.

Профиль (направленность) программы 05.13.18 Математическое моделирование, численные методы и комплексы программ

Аспирантская школа по компьютерным наукам

Аспирант _____/ Гречихин И.С. /
подпись

Научный руководитель _____/Колданов А.П. /
подпись

Директор Аспирантской школы по компьютерным наукам_____/ Объедков С.А./
подпись

Нижний Новгород, 2020

Тема диссертации: Статистическая неопределённость алгоритмов идентификации графических моделей.

Актуальность исследования

Исследованию графических моделей привлекает активное внимание исследователей последние десятилетия. Сетевые модели помогают решать задачи связанные с анализом больших массивов данных. Такие модели имеют приложения в биоинформатике, поиске информации в текстах и в современных кодах с исправлением ошибок: многие, практически важные сетевые модели (сеть экспрессии генов, сеть коэкспрессии генов, сеть взаимодействия нейронов головного мозга, сеть активов фондового рынка и др.) могут быть представлены как сети случайных величин (random variable network). Задача идентификации состоит в восстановлении графической модели по наблюдениям. В существующих исследованиях основное внимание уделяется алгоритмам идентификации и их применению для решения различных задач. При этом, как правило, без внимания остаётся вопрос достоверности получаемых выводов.

Цель исследования

Исследование статистической неопределённости алгоритмов идентификации графических моделей.

Задачи исследования

- Обзор существующих вычислительных алгоритмов и методов оценки их статистической неопределённости.
- Развитие нового подхода к оценке статистической неопределённости процедур идентификации графических моделей.
- Разработка программного модуля для сравнения статистической неопределённости известных процедур идентификации графических моделей.
- Проведение вычислительных экспериментов для сравнения статистической неопределённости известных процедур идентификации графических моделей.

Анализ современного состояния исследований в данной области

Сетевой подход к анализу сложных систем и больших массивов данных вызывает большой интерес в последние годы. Это отражено в растущем числе публикаций на эту тему, включая ряд монографий и обзоров (Cowell at al. 1999), (Carrington at al. 2005), (Koller & Friedman 2009), (Horvath 2011), (Wainwright & Jordan, 2008). Значительное место в этих исследованиях занимают вероятностные графические (графовые) модели. Вероятностная графическая модель представляет структуру зависимостей семейства случайных величин в терминах ориентированных и неориентированных графов (Edwards, 2000), (Jordan, 2004). При этом вершины графа идентифицируются со случайными величинами, а ребра графа отражают характер зависимостей этих случайных величин (Cox & Wermuth, 1993), (Cox & Wermuth, 1996).

Проблема идентификации гауссовской графической модели заключается в построении статистических процедур выбора графической модели по наблюдениям случайных величин с совместным нормальным распределением. Одним из теоретически обоснованных подходов к идентификации гауссовских графических моделей является байесовский подход. В рамках этого подхода разработаны различные алгоритмы идентификации (Schäfer and Strimmer, 2005). Другой фундаментальный подход связан с использованием индивидуальных тестов проверки гипотез об условных независимостях, определяющих модель (Wermuth, 1976), (Drton and Perlman, 2004).

В последние годы начато развитие этого подхода, основанное на применении процедур одновременной проверки многих гипотез (Hochberg and Tamhane, 1987). Основной целью при этом является построение процедур, контролирующих вероятность хотя бы одного ложного отвержения индивидуальной гипотезы при любом количестве истинных индивидуальных гипотез (Lehmann and Romano, 2005). Развитие этой теории привело к построению одношаговых и многошаговых процедур одновременной проверки многих гипотез для идентификации гауссовских графических моделей. В работе (Drton and Perlman, 2004) рассматривается многомерное нормальное распределение и изучается

одношаговая процедура построения гауссовской графической модели, основанная на построении доверительных интервалов для частных коэффициентов корреляции. В работе (Drton and Perlman, 2007) исследуется многошаговая процедура одновременной проверки многих гипотез о равенстве нулю частного коэффициента корреляции. Для контроля FWER используется процедура Холма (Holm, 1979). В работах (Drton and Perlman, 2007), (Cai and Liu, 2016) изучаются многошаговые статистические процедуры, контролирующие FDR (математическое ожидание доли ложных отвержений), FDP (вероятность того, что доля ложных отвержений больше заданного порога). Полученные результаты, в основном, ограничены исследованием вероятностных графических моделей с многомерным нормальным распределением.

Вместе с тем, в задачах идентификации графовых (графических) моделей естественно рассматривать более общие распределения, такие как экспоненциальный класс распределений, и контролировать различные характеристики качества идентификации, в том числе используемые в близких задачах бинарной классификации. Анализ статистических процедур идентификации графических моделей с этих позиций только начал развиваться в последние годы.

Объект исследования

Гауссовские графические модели, алгоритмы идентификации гауссовских графических моделей.

Методы исследования

Методы теории вероятности и математической статистики, вычислительной математики, методы оптимизации и другие методы компьютерных наук. Инструменты программирования.

Полученные результаты

- Рассмотрены:
    - различные процедуры идентификации графических моделей: GGM (использование полных частных корреляций), 0-1 (частные корреляции 1го порядка), 0-1-2 (частные корреляции 2го порядка);
    - различные типы поправок к корреляционным коэффициентам;
    - меры ошибок (FWER, FDR, ошибки 1го, 2го рода, функции риска)
- Создан набор программ на Python для сравнения различных процедур
- Проведены вычислительные эксперименты для графовых моделей разного размера и различных предположений о распределении данных
- Полученные сравнения говорят о нестабильности процедур при изменении распределения

Доклады и конференции

- NET 2017: Участие с докладом «Comparison of different methods of graphical models identification» (доклад попал в сборник)
- XI Международная конференция «Применение многомерного статистического анализа в экономике и оценке качества»: участие с докладом «Устойчивость статистических процедур построения графовых моделей»

Публикации

- Grechikhin I., Kalyagin V. A. (2018) Comparison of statistical procedures for Gaussian graphical model selection, in: Computational Aspects and Applications in Large-Scale Networks. Springer Proceedings in Mathematics & Statistics Vol. 247. Springer. P. 269-279.
- Grechikhin I., Kalyagin V., Koldanov A. On a robustness of statistical procedures for graphical model selection, тезисы доклада на XI-ой Международной научной конференции «Применение многомерного статистического анализа в экономике и оценке качества», Москва, 21-23 августа 2018г.

Участие в научных проектах

- грант РФФИ 18-07-00524 "Методы принятия решений в задачах идентификации графовых моделей" в 2018-2020гг.
- участие в проектах лаборатории Алгоритмов и технологий анализа сетевых структур НИУ ВШЭ,
- участие в НУГ «Анализ мультимедийных данных пользователей мобильных устройств»

Прикладные исследования по близкой тематике

В ходе участия в НУГ «Анализ мультимедийных данных пользователей мобильных устройств» я работал над задачами связанными с классификацией и детектированием объектов на изображениях. Задача классификации связана с исследованием статистической неопределённости, поскольку наличие/отсутствие связей в графических моделях можно решать с помощью алгоритмов классификации; статистическая неопределённость в этом случае становится мерой уверенности в ответе алгоритма классификации.

В ходе работ решалась задача детектирования объектов ~70 различных категорий на изображениях, полученных с помощью мобильных устройств, а также классификация субкатегорий некоторых выделенных категорий. Целью работы являлось получение предпочтений пользователя на основе его мультимедийных данных.

Доклады и конференции:

IbPRIA 2019: Постерный доклад “User Modeling on Mobile Device based on Facial Clustering and Object Detection in Photos and Videos”

MMPO 2019: Доклад “User Modeling on Mobile Device based on Facial Clustering and Object Detection in Photos and Videos”

Публикации:

- И. С. Гречихин, А. В. Савченко. (2019) Метод анализа предпочтений пользователя по фото- и видеоизображениям на мобильном устройстве на

основе нейросетевых детекторов объектов на изображениях. Информационные технологии. Т. 25. № 9. С. 538-544

- Grechikhin I., Andrey V. Savchenko. (2019) User Modeling on Mobile Device Based on Facial Clustering and Object Detection in Photos and Videos, in: Pattern Recognition and Image Analysis Part 2. Springer. P. 429-440
- А. В. Савченко, И. С. Гречихин. (2020) Детектирование специализированных категорий объектов на фотографиях в мобильных устройствах на основе многозадачной нейросетевой модели. Информационные технологии. Т. 26. № 10. С. 586-593.

Литература

1. Anderson T.W. (2003) An introduction to multivariate statistical analysis. third edition, Wiley-Interscience, New-York.
2. Beerwinkel, N. and Drton, M. (2007). A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data. Biostatistics 8 53–71.
3. Cai T.T. and Liu W. (2016) Large-scale multiple testing of correlations. Journal of the American Statistical Association, 513(111):229-240.
4. Carrington P.J., Scott J., Wasserman S. (2005) Models and methods in social network analysis, Cambridge university press.
5. Cooper, G. F. and Herskowits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9 309–347.
6. Cowell, R. G., David, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999). Probabilistic Networks and Expert Systems. Springer, New York.
7. Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. Statist. Sci. 8 204–218. MR1243593
8. Cox, D. R. and Wermuth, N. (1996). Multivariate Dependencies. Chapman and Hall, London.
9. Dellaportas, P., Giudici, P. and Roberts, G. (2003). Bayesian inference for nondecomposable graphical Gaussian models. Sankhya 65 43–55.
10. Dempster A.P. (1972) Covariance selection, Biometrics28, 157-175.
11. Drton M. and Perlman M. (2004) Model selection to Gaussian concentration graph. Biometrika, 91(3), 591-602.
12. Drton M. and Perlman M. (2007) Multiple testing and error control in gaussian graphical model selection. Statistical Science, 22(3): 430-449.

13. Drton M. Maathuis M.H. (2017) Structure Learning in Graphical Modeling, Annual Review of Statistics and Its Applications, v.4, pp. 365-393.
14. Edwards D.M. (2000) Introduction to graphical modeling, Springer, NY, 2-d edition.
15. Fang K.T. Kotz S. Ng K.W. (1990) Symmetric multivariate and related distributions. Chapman and Hall, London.
16. Gupta F.K., Varga T., Bodnar T. (2013) Elliptically Contoured Models in Statistics and Portfolio Theory. Springer.
17. Hochberg Y. and Tamhane A. (1987) Multiple comparison procedures. John Wiley& Sons, Inc.
18. Holm S. (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2):65-70.
19. Horvath, S. (2011) Weighted Network Analysis. Applications in Genomics and Systems Biology, Springer Book, ISBN 978-1-4419-8818-8.
20. Jordan, M. I. (2004). Graphical models. Statist. Sci. 19 140–155.
21. Lehmann E. (1957) A theory of some multiple decision problems I. The Annals of Mathematical Statistics, 1-25.
22. Lehmann E.L.and Romano J.P. (2005) Testing statistical hypotheses. Springer, New York.
23. Schäfer J. and Strimmer K. (2005) An empirical Bayes approach to inferring large-scale gene association networks, BIOINFORMATICS, Vol. 21 no. 6 , pp. 754–764.
24. Schwartz, G. (1978). Estimating the dimension of a model. Ann.Statist. 6 461–464.
25. Sidak Z.(1967) Rectangular con_dence regions for the means of multivariate normal distributions. J. Am. Statist. Assoc., 62:626-633.
26. Wainwright M.J. Jordan M.I. (2008) Graphical Models, Exponential Families, and Variational Inference, Foundations and Trends in Machine Learning Vol. 1, Nos. 1–2, 1–305.
27. Wei Gao Wenna Ye. (2018) A Min-Max conditional covariance algorithm for structure learning of Gaussian graphical models, Stat Anal Data Min: The ASAData Sci Journal;1–11.
28. Wenbin Guo, Cristiane P. G. Calixto, Nikoleta Tzioutziou, Ping Lin, Robbie Waugh, John W. S. Brown and Runxuan Zhang. (2017) Evaluation and improvement of the regulatory inference for large co-expression networks with limited sample size, BMC Systems Biology 11:62.
29. Wermuth, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. Biometrics 32, 95–108.
30. Xingqi Du, Subhashis Ghosa. (2019) Multivariate Gaussian network structure learning, Journal of Statistical Planning and Inference, v. 199, p. 327–342.
31. Zuo Y, Yu G, Tadesse MG, Ressom HW. (2014) Biological network inference using low order partial correlation. Methods (San Diego, Calif). v.69, pp. 266–273.