ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐẶNG THỊ LỆ UYÊN

Lựa chọn mô hình với dữ liệu khuyết nhiều chiều bằng phương pháp ADAPTIVE BAYESIAN SLOPE

NGÀNH CƠ SỞ TOÁN CHO TIN HỌC
CHUYÊN NGÀNH KHOA HỌC DỮ LIỆU

Giảng viên hướng dẫn: **TS. Hoàng Văn Hà**

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐẶNG THỊ LỆ UYÊN
Mã số học viên: 21C29027

ĐỀ CƯƠNG NGHIÊN CỨU ĐỀ TÀI LUẬN VĂN THẠC SĨ

Tên đề tài

Lựa chọn mô hình với dữ liệu khuyết nhiều chiều bằng phương pháp ADAPTIVE BAYESIAN SLOPE

Ngành CƠ SỞ TOÁN CHO TIN HỌC
Chuyên ngành KHOA HỌC DỮ LIỆU
Mã số chuyên ngành: 8460108

Xác nhận của giảng viên hướng dẫn:
(Ký tên và ghi rõ họ tên)

Họ tên: .....

TPHCM, tháng 4 năm 2023

1 Giới thiệu tổng quan

Giới thiệu tổng quan

Việc lựa chọn các biến từ dữ liệu nhiều chiều là một bài toán phổ biến trong nhiều ứng dụng dữ liệu hiện đại. Ví dụ, trong di truyền học phân tử, có rất nhiều biến dự báo nhưng chỉ một số ít được coi là phù hợp để giải thích các hiện tượng sinh học. LASSO (Tibshirani, 1996), là một phương pháp phạt phổ biến nhất hiện nay, đã thành công trong việc ước lượng các tham số và đồng thời thực hiện "lựa chọn biến" trong mô hình. Mặc dù LASSO sở hữu những đảm bảo tốt về mặt lý thuyết, nhưng nó có thể dẫn đến "phát hiện sai lầm" (*false discoveries* - (Su và cộng sự, 2017) và nó chỉ cho phép xác định mô hình trong các điều kiện "không thể đại diện" khá nghiêm ngặt (Wainwright, 2009; Tardivel và Bogdan 2018). Một biến thể của LASSO - adaptive LASSO (Zou, 2006), sử dụng hàm phạt có trọng số l_1 để làm giảm sai lệch trong việc ước lượng tham số và có thể nhất quán trong việc lựa chọn biến ngay cả khi điều kiện "không thể đại diện" không thỏa (Fan và Barut 2014; Tardivel và Bogdan 2018; Tardivel và Bogdan 2019). Tuy nhiên, các tính chất về mặt hiệu suất của adaptive LASSO vẫn chủ yếu dựa vào hàm trọng số và các tham số hiệu chỉnh mà sự lựa chọn tối ưu phụ thuộc vào các khía cạnh chưa biết của bài toán ước lượng như là cường độ hoặc độ thưa của tín hiệu.

Gần đây, Ročková và George (2018) đã phát triển quy trình Spike-và-Slab LASSO (SSL), quy trình này kết hợp phương pháp phạt mặc định (LASSO) và phương pháp chọn biến Bayes mặc định (spike-và-slab). Trong SSL, hàm phạt phát sinh từ công thức Bayes spike-and-slab đầy đủ và do đó, tạo ra các tính chất tự thích ứng (self-adaptation) ít phải hiệu chỉnh siêu tham số hơn. Ngoài ra, SSL giảm bớt sự co gọn quá mức của các biến quan trọng bằng cách cung cấp hỗ trợ tiên nghiệm đầy đủ cho chúng. Kết quả lý thuyết và mô phỏng trong bài báo của Ročková và George (2018) cho thấy SSL đạt được tốc độ hội tụ gần minimax (cho cả chế độ hậu nghiệm và hậu nghiệm toàn phần) và hiệu suất rất tốt ngay cả khi các cột trong ma trận thiết kế (design matrix) có tương quan mạnh với nhau.

Bogdan và cộng sự (2015) đã đề xuất phương pháp SLOPE (Sorted L-One Penalized Estimator) để kiểm soát tỷ lệ phát hiện sai (False Discovery Rate - FDR). SLOPE kiểm soát FDR khi ma trận thiết kế là trực giao. Hơn nữa, Su và Candès (2016) và Bellec và cộng sự (2018) đã cho thấy rằng, trái ngược với LASSO, SLOPE cho phép mô hình đạt được tốc độ hội tụ minimax chính xác đối với các hệ số hồi quy trong hồi quy nhiều chiều thưa thớt. Tuy nhiên, tương tự như LASSO, để vừa đạt được dự đoán tốt vừa chọn được biến tốt trong mẫu hữu hạn là một thách thức với SLOPE. Để FDR nhỏ thì cần một lượng lớn co gọn, điều này dẫn đến sai lệch

lớn trong việc ước lượng các hệ số hồi quy quan trọng, và do đó ước lượng kém. Bogdan và cộng sự (2015); Brzyski và cộng sự (2019) đã đề xuất một biện pháp khắc phục gồm hai bước như sau: Bước 1) sử dụng SLOPE để phát hiện các biến độc lập có liên quan; Bước 2) áp dụng bình phương bé nhất tiêu chuẩn (standard least squares) với các biến được chọn trong bước 1 để ước lượng. Cách tiếp cận này cho phép mô hình giảm bớt sai lệch của SLOPE. Tuy nhiên, khi các cột trong ma trận thiết kế tương quan với nhau, vấn đề mất kiểm soát FDR vẫn còn tồn tại. Việc mất kiểm soát FDR này là kết quả của việc co gọn quá mức các hệ số hồi quy lớn, mà ảnh hưởng không giải thích được của chúng thường được bù đắp bằng các biến giải thích "giả" tương quan nhẹ (Su và các cộng sự (2017)).

Jiang và cộng sự (2019) đã đề xuất Adaptive Bayesian SLOPE (ABSLOPE) - nhúng SLOPE vào bộ khung Bayesian spike-và-slab. Trong mô hình này, phần tiên nghiệm được xây dựng sao cho thành phần "spike" giảm thành SLOPE thông thường đối với các hệ số hồi quy rất nhỏ. Đồng thời với việc giảm sai lệch của các tín hiệu quan trọng bằng thành phần "slab" cho phép kiểm soát FDR trong nhiều tình huống có thể xảy ra. Hơn nữa, thành phần "slab" của tiên nghiệm hỗn hợp (mixture prior) bảo toàn tính chất trung bình của SLOPE đối với các hệ số hồi quy tương tự nhau (Figueiredo và Nowak, 2016). Điều này dẫn đến chất lượng dự đoán rất tốt khi các biến hồi quy có mối tương quan mạnh. Các siêu tham số của tiên nghiệm SLOPE hỗn hợp được cập nhật lặp đi lặp lại bằng cách sử dụng mô hình Bayes đầy đủ với thuật toán xấp xỉ ngẫu nhiên EM (Lavielle, 2014), thuật toán này cũng có thể xử lý dữ liệu khuyết. Do đó, mục tiêu của ABSLOPE là chọn biến với dữ liệu nhiều chiều và bị khuyết.

Xử lý dữ liệu khuyết trong bối cảnh lựa chọn biến nhiều chiều là một vấn đề rất quan trọng. Thật vậy, dữ liệu khuyết ở khắp mọi nơi, có nhiều lý do dẫn đến việc dữ liệu bị khuyết và các cách xử lý dữ liệu khuyết thông thường, như xóa bỏ, dẫn đến sự sai lệch nếu dữ liệu bị khuyết không hoàn toàn ngẫu nhiên, và mất mát thông tin. Không thiếu các tài liệu về việc quản lý dữ liệu khuyết nhưng chỉ có vài phương pháp hỗ trợ chọn mô hình khi dữ liệu bị khuyết. Ví dụ, trong các mô hình tuyến tính tổng quát, Claeskens và Consentino (2008); Ibrahim và cộng sự (2008); Jiang và cộng sự (2019) đã điều chỉnh các tiêu chí thông tin (information criteria) dựa trên hàm hợp lý được thiết kế cho dữ liệu đầy đủ như AIC. Tuy nhiên, các phương pháp đó không thể xử lý dữ liệu lớn khi số chiều p lớn hơn cỡ mẫu n. Trong các mô hình tuyến tính, Loh và Wainwright (2012) đã xây dựng một biến thể LASSO bằng cách sửa đổi ước lượng ma trận hiệp phương sai trong trường hợp dữ liệu bị khuyết, và đã giải quyết bài toán không lồi (non convex problem) với một thuật toán dựa vào projected gradient descent. Tuy nhiên, phương pháp này giả định rằng chuẩn l_1

bị giới hạn bởi một hằng số và hằng số này phụ thuộc vào độ thưa thớt hiếm gặp trong thực tế. Trong bài báo khác, Zhao và cộng sự (2017) đã đề xuất phương pháp pseudo-likelihood với hàm phạt LASSO, được sử dụng chọn biến nhưng không ước lượng tham số. Cuối cùng, Liu và cộng sự (2016) đã kết hợp các kỹ thuật hồi quy phạt để điền khuyết nhiều lần (multiple imputation) và lựa chọn biến.

2 Mục đích nghiên cứu

Luận văn sẽ đưa ra lý thuyết thống kê cần thiết, các phương pháp xử lý dữ liệu khuyết cơ bản. So sánh phương pháp ABSLOPE được đề xuất với các phương pháp trước đó trên dữ liệu mô phỏng và dữ liệu thực tế.

Luận văn này dựa trên bài báo *"Adaptive Bayesian SLOPE - High-dimensional Model Selection with Missing Values"* của các tác giả Wei Jiang, Malgorzata Bogdan, Julie Josse, Blažej Miasojedow, Veronika Ročková, TraumaBase® Group.

3 Đối tượng nghiên cứu

Luận văn hướng đến các đối tượng: các phương pháp ước lượng tham số, lựa chọn các biến đối với dữ liệu nhiều chiều và bị khuyết.

4 Các phương pháp nghiên cứu

Phương pháp khoa học: đọc hiểu tài liệu, hệ thống kiến thức, kiểm chứng lý thuyết.

5 Nội dung và phạm vi của vấn đề sẽ đi sâu nghiên cứu

Nội dung của luận văn bao gồm 5 chương:

Chương 1: Kiến thức chuẩn bị

1.1 LASSO

1.2 SLOPE

1.3 Adaptive Bayesian SLOPE

Chương 2: Dữ liệu khuyết

2.1 Tổng quan về dữ liệu khuyết

2.2 Các phương pháp xử lý dữ liệu khuyết

Chương 3: Ước lượng tham số và chọn mô hình

3.1 Thuật toán EM

3.2 Thuật toán SAEM

3.3 SLOBE

Chương 4: Mô phỏng nghiên cứu

4.1 Cài đặt mô phỏng

4.2 So sánh ABSLOPE và SLOBE với các phương pháp khác

Chương 5: Ứng dụng trên dữ liệu thực

5.1 Giới thiệu tập dữ liệu

5.2 Kết quả đạt được

6 Nơi thực hiện đề tài nghiên cứu

Đề tài luận văn thạc sĩ được thực hiện tại trường Đại học Khoa học Tự nhiên - Đại học Quốc gia Thành Phố Hồ Chí Minh.

7 Thời gian thực hiện

Luận văn được thực hiện từ tháng 5 năm 2023 đến tháng 11 năm 2023.

Tài liệu tham khảo

Bellec, P., L. G. and Tysbakov, A. (2018). Slope meets Lasso: improved oracle bounds and optimality. *Ann. Statist.*, 46(6B):3603–3642.

Bogdan, M., v. d. B. E. S. C. S. W. and Candès, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140.

Brzyski, D., G. A. S. W. and Bogdan, M. (2019). Group SLOPE – adaptive selection of groups of predictors. *Journal of the American Statistical Association*, 114(525):419–433.

Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, 64:1062–9.

Fan, J., F. Y. and Barut, E. (2014). Adaptive robust variable selection. *Annals of Statistics*, 42(1):324–351.

Figueiredo, M. A. T. and Nowak, R. D. (2016). Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, JMLR:WCP*, 51:930–938.

Ibrahim, J., Z. H. and Tang, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association*, 103(484):1648–1658.

Jiang, W., B. M. J. J. M. B. R. V. and Group, T. (2019). Additional supplementary materials for Adaptive Bayesian SLOPE – high-dimensional model selection with missing values.

Lavielle, M. (2014). Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools. *Chapman and Hall/CRC*.

Liu, Y., W. Y. F. Y. and Wall, M. M. (2016). Variable selection and prediction with incomplete high-dimensional data. *Ann. Appl. Stat.*, 10(1):418–450.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664.

Rejchel, W. and Bogdan, M. (2019). Fast and robust model selection based on ranks. *arXiv preprint*, 1905.05876.

Ročková, V. and George, E. I. (2018). The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 113(521):431–444.

Su, W. and Candès, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Appl. Stat.*, 44(3):1038–1068.

Su et al., W., B. M. C. E. (2017). False discoveries occur early on the Lasso path. *The Annals of Statistics*, 45(5):2133–2150.

Tardivel, P. J. and Bogdan, M. (2018). On the sign recovery given by the thresholded LASSO and thresholded Basis Pursuit. *arXiv preprint*, arXiv:11812.05723.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1).

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming (Lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Zhao, J., Y. Y. and Ning, Y. (2017). Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data, 28.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.