

QT32: READING PAPER #2

Generalized probabilistic principal component analysis of correlated data

<https://www.jmlr.org/papers/volume21/18-595/18-595.pdf>

Thành viên nhóm: Trần Bình Hậu – CH2001004

Nguyễn Thanh Phong – CH2001012

Trần Văn San – CH2001013

Generalized probabilistic principal component analysis of correlated data

Mengyang Gu

MENGYANG@PSTAT.UCSB.EDU

*Department of Statistics and Applied Probability
University of California, Santa Barbara
5511 South Hall
Santa Barbara, CA 93106-3110*

Weining Shen

WEININGS@UCI.EDU

*Department of Statistics
University of California, Irvine
2206 Bren Hall
Irvine, CA 92697-1250*

Editor: Manfred Opper

PLEASE NOTE: ARC no longer uses the ERA2010 rankings list. The 2020 ranking is the result of a partial evaluation, reviewing only journals that were identified by metrics as being possibly misranked. A future round will allow for additions and community initiated reviews.
For further details of the 2020 process please see the CORE website

[Sign in with LinkedIn](#)

Signing in with LinkedIn authorizes us to store your name, email address, headline and display picture

[why?](#)

A* - 6%

A - 10%







B - 30%

C - 43%

Other - 8%

Journal of Machine Learning Research Search by: Source:

Showing results 1 - 1 of 1

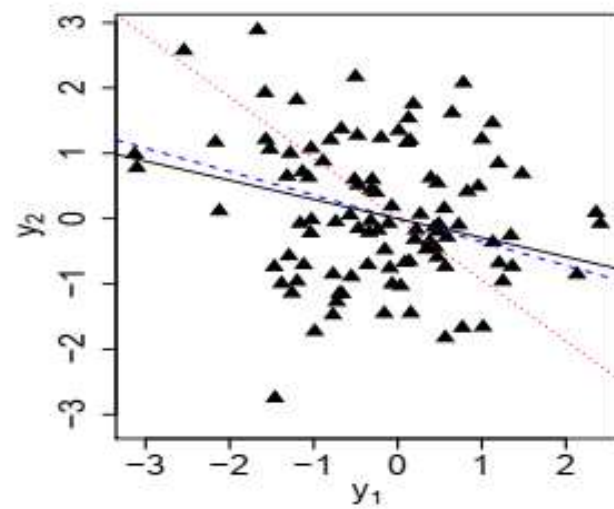
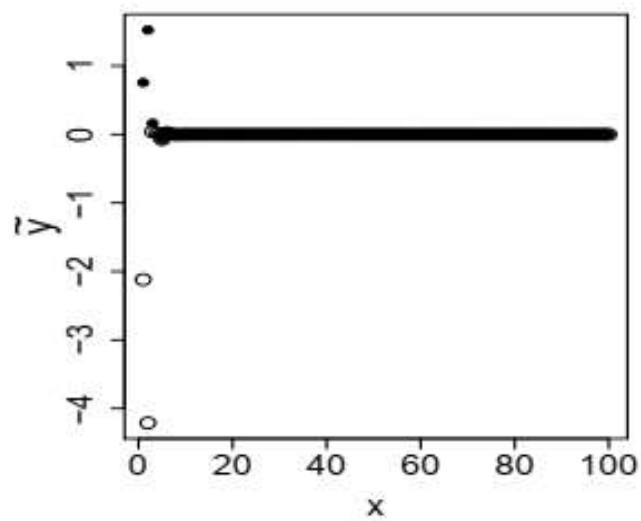
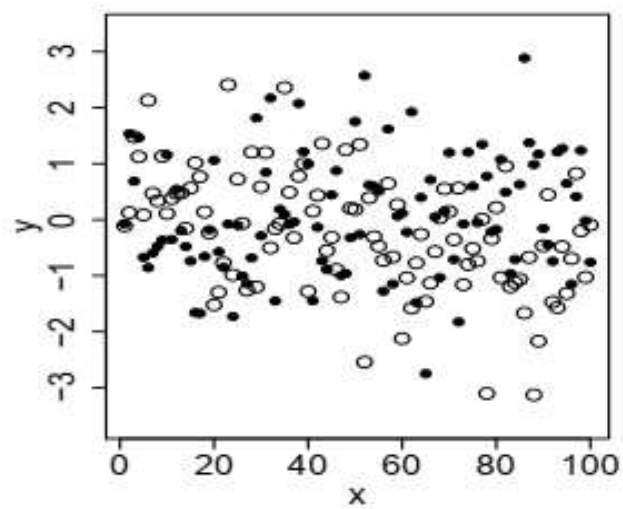
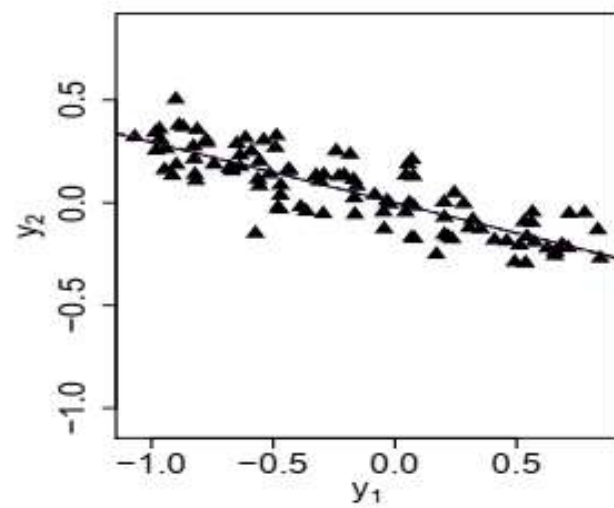
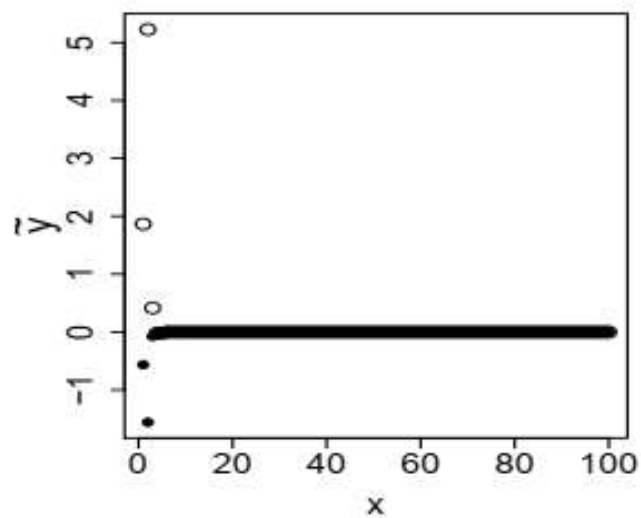
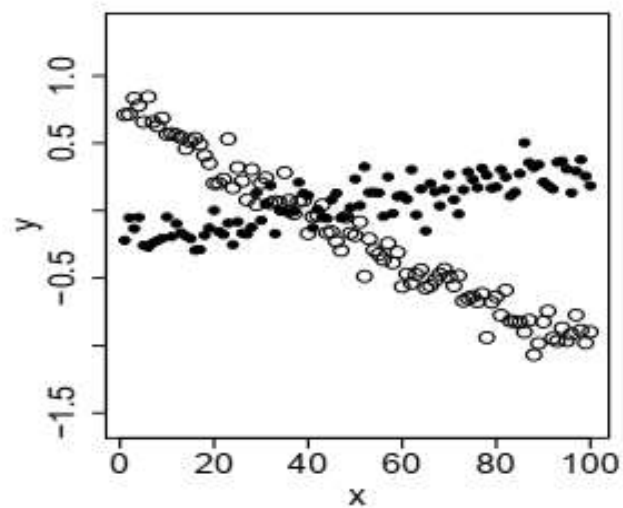
Title 	Source 	Rank 	DBLP 	Has data? 	FoR 	Comments 	Average Rating 
Journal of Machine Learning Research	CORE2020	A*	view	Yes	0801	4	5.0

1-Bài toán mà bài báo giải quyết là gì? Minh hoạ input/output (tìm hình ảnh có trong bài báo để minh hoạ).

- Phân tích thành phần chính (PCA) là một trong những kỹ thuật quan trọng trong học máy và xử lý dữ liệu nhằm giảm số chiều dữ liệu.

Input: Véc tơ $x = (x_1, x_2, \dots, x_p)^T$

Output: Véc tơ $y(x) = (y_1(x), y_2(x), \dots, y_k(x))^T$ với $k < p$.



2- Các câu hỏi đặt ra là gì? Đã giải quyết được đến đâu?

- Sử dụng phương pháp ước lượng hợp lý cực đại ta có các thành phần chính tương đương với các giá trị riêng của ma trận tương quan của véc tơ dữ liệu quan sát và giả định rằng các thành phần của véc tơ dữ liệu **có phân phối độc lập theo phân phối chuẩn và không tương quan**. Tuy nhiên, giả định về tính độc lập có thể không thực tế đối với nhiều trường hợp như mô hình hóa nhiều chuỗi thời gian, quá trình không gian (spatial processes) và hàm dữ liệu, trong đó các thành phần véc tơ dữ liệu có sự tương quan.
- Trong bài báo này, tác giả giới thiệu phương pháp xác suất tổng quát phân tích thành phần chính (GPPCA) để nghiên cứu **mô hình dữ liệu cho nhiều thành phần tương quan**, trong đó mỗi thành phần được mô hình hóa bằng một quá trình Gaussian. Phương pháp này đã **đưa ra công thức tổng quát hơn để xác định các tham số** so với trước đây bằng phương pháp ước lượng hợp lý cực đại.

- Dựa vào ma trận biểu thị độ chính xác trong ước lượng hợp lý cực đại đưa ra số lượng các biến tính toán là tuyến tính với số lượng biến đầu ra.
- Hơn nữa, nhóm tác giả đưa ra biểu thức của ước lượng hợp lý cực đại cho kỳ vọng cho các biến tương quan.
- Lợi thế của GPPCA xét về mức độ phù hợp thực tế, ước tính độ chính xác và tính thuận tiện trong tính toán. Thí nghiệm trên bộ dữ liệu thực và dữ liệu mô phỏng đạt được hiệu quả cao hơn so với các cách tiếp cận khác

3- Ý tưởng giải quyết là gì? Minh họa trực quan (diễn giải bằng lời hoặc hình ảnh)

Ý tưởng chính: Dùng thuật toán gradient descent để ước lượng hợp lý cực đại của ma trận \mathbf{A} trong mô hình $y(\mathbf{x}) = \mathbf{A}\mathbf{z}(\mathbf{x}) + \epsilon$.

To begin with, let $\mathbf{y}(\mathbf{x}) = (y_1(\mathbf{x}), \dots, y_k(\mathbf{x}))^T$ be a k -dimensional real-valued output vector at a p -dimensional input vector \mathbf{x} . Let $\mathbf{Y} = [\mathbf{y}(\mathbf{x}_1), \dots, \mathbf{y}(\mathbf{x}_n)]$ be a $k \times n$ matrix of the observations at inputs $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. In this subsection and the next subsection, we assume that each row of the \mathbf{Y} is centered at zero.

Consider the following latent factor model

$$\mathbf{y}(\mathbf{x}) = \mathbf{A}\mathbf{z}(\mathbf{x}) + \epsilon, \quad (1)$$

where $\epsilon \sim N(0, \sigma_0^2 \mathbf{I}_k)$ is a vector of independent Gaussian noises, with \mathbf{I}_k being the $k \times k$ identity matrix. The $k \times d$ factor loading matrix $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_d]$ relates the k -dimensional output to d -dimensional factor processes $\mathbf{z}(\mathbf{x}) = (z_1(\mathbf{x}), \dots, z_d(\mathbf{x}))^T$, where $d \leq k$.

Theorem 3 *Under Assumption 1, after marginalizing out \mathbf{Z} , the maximum marginal likelihood estimator of \mathbf{A} in model (1) is*

$$\hat{\mathbf{A}} = \underset{\mathbf{A}}{\operatorname{argmax}} \sum_{l=1}^d \mathbf{a}_l^T \mathbf{G}_l \mathbf{a}_l, \quad s.t. \quad \mathbf{A}^T \mathbf{A} = \mathbf{I}_d, \quad (7)$$

where $\mathbf{G}_l = \mathbf{Y}(\sigma_0^2 \mathbf{\Sigma}_l^{-1} + \mathbf{I}_n)^{-1} \mathbf{Y}^T$.