# 10.7 Latent Variable Perspective

Phần này nói về một cách nhìn khác về PCA (Principal Component Analysis), thay vì chỉ xem PCA là một phương pháp toán học để giảm chiều dữ liệu, ta có thể nhìn nó dưới góc độ mô hình xác suất.

PPCA (Probabilistic PCA) là phiên bản xác suất của PCA, được giới thiệu bởi Tipping và Bishop năm 1999. Trong PPCA:

* Ta giả định rằng dữ liệu được tạo ra từ một biến ẩn liên tục z (có phân phối chuẩn).
* Dữ liệu quan sát là một phiên bản bị nhiễu của không gian biến ẩn đó.
* Khi dùng ước lượng hợp lý cực đại (maximum likelihood estimation) trong trường hợp không nhiễu, kết quả thu được chính là PCA cổ điển.

Mục tiêu của PPCA là biến PCA từ một công cụ toán học thành một mô hình sinh dữ liệu xác suất (probabilistic generative model) bằng cách đưa vào biến ẩn (latent variable) z.

## Ý tưởng

Trong PPCA, chúng ta giả định rằng những dữ liệu mà quan sát được thực chất được sinh ra từ một nguyên nhân ẩn sâu bên trong gọi là biến ẩn (latent variable) z. Vậy nên, trong mô hình PPCA (Probabilistic PCA), ta giả định rằng dữ liệu được sinh ra theo công thức:

với các giả định sau:

* Biến latent **:** (đối với , ).
* Nhiễu **:** (nhiễu Gaussian với phương sai ).

Như vậy, điều kiện cho , tức là , có dạng giống như được tạo ra bằng cách cộng một nhiễu Gaussian vào một vector μ + Bz:

Tiếp theo ta cần tính phân phối biên của x bằng cách “tích phân loại bỏ” z (vì z và đều là Gaussian), Ta cần tính:

Vì , và , và độc lập với nhau, nên:

Theo định lý về tổng của các biến Gaussian, tổng hai Gaussian độc lập vẫn là Gaussian, với:

* Kỳ vọng (mean):
* Hiệp phương sai (covariance):

Vì:

Tổng lại:

Từ đó, ta có thể suy ra phân phối biên của :

Như vậy, mô hình PPCA cho rằng ma trận hiệp phương sai của dữ liệu là

## Mối liên hệ với phân rã giá trị riêng (Eigen-decomposition)

Ta có một tập dữ liệu gồm N mẫu, cần tìm :

Sao cho likelihood (hoặc log-likelihood) là lớn nhất.

Ta ước lượng covariance mẫu từ tập dữ liệu là . Theo mô hình trên, ta mong muốn:

Một cách tự nhiên để giải phương trình này là sử dụng phân rã giá trị riêng (eigendecomposition) của . Cụ thể:

Phân rã , ta có thể viết:

trong đó là ma trận các vector riêng (các cột là các vector riêng) và là ma trận đường chéo với các giá trị riêng . Các vector riêng này cho ta "hướng chính" của phương sai dữ liệu; giá trị riêng tương ứng cho biết lượng phương sai dọc theo hướng đó.

Trong không gian latent chiều, thông tin chính được giữ lại ở các hướng có giá trị riêng lớn nhất, còn phần còn lại chỉ là nhiễu (với mức ). Cụ thể, nếu ta chọn thành phần chính, ta có thể viết giống như phần "giải thích" được bởi thành phần đó.  
Theo kết quả của Tipping và Bishop, nghiệm của bài toán maximum likelihood cho PPCA có dạng:

với:

* là ma trận chứa vector riêng ứng với giá trị riêng lớn nhất.
* là ma trận đường chéo chứa giá trị riêng này.
* là một ma trận quay (rotation matrix) tùy ý, có thể chọn (trung hòa được bước không xác định này).

Vì vậy, nếu chọn , ta được công thức:

## Ước lượng

Còn lại là ước lượng nhiễu , trong mô hình PPCA, các thành phần không thuộc không gian latent ( đến ) được cho là chỉ chứa nhiễu, tức là:

Một cách ước lượng đơn giản là lấy trung bình các giá trị riêng này:

## Sinh dữ liệu mới

Để sinh ra dữ liệu mới với những tham số được cho, chúng ta làm theo một quy trình gọi là Ancestral Sampling:

1. Lấy mẫu z từ:
2. Sinh ra x bằng cách đưa z qua ánh xạ

Cụ thể, sau khi có và ước lượng , ta có thể sinh mẫu mới bằng cách:

Trong đó:

* với kích thước (sinh mẫu từ không gian latent).
* Nếu muốn thêm nhiễu, ta thêm

## Tham khảo

[**https://www.cs.columbia.edu/~blei/seminar/2020-representation/readings/TippingBishop1999.pdf**](https://www.cs.columbia.edu/~blei/seminar/2020-representation/readings/TippingBishop1999.pdf)