BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC QUY NHƠN

NGUYỄN NHẬT NAM

MỘT SỐ PHƯƠNG PHÁP PHÂN TÍCH DỮ LIỆU VÀ ỨNG DỤNG

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC
Ngành: TOÁN ỨNG DỤNG

Người hướng dẫn
TS.LÂM THỊ THANH TÂM

Bình Định, 2023

MỞ ĐẦU

Nội dung khóa luận được trình bày trong 4 chương:

- **Chương 1. Một số kiến thức chuẩn bị**
- **Chương 2. Phân tích giá trị kì dị (SVD)**
- **Chương 3. Phân tích thành phần chính (PCA)**
- **Chương 4. Một số ứng dụng của SVD và PCA**

- **Định lí về sự tồn tại của SVD**
- **Thuật toán SVD**

Định lý

Cho $\mathbf{A}$ là một ma trận có cấp $m \times n$. Khi đó, mọi giá trị riêng của ma trận $\mathbf{A}^T\mathbf{A}$ đều không âm.

Định lý

Mọi ma trận $\mathbf{A}$ cỡ $m \times n$ bất kì đều có phân tích SVD có dạng

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T.$$

Định lý

Cho $\mathbf{A}$ là ma trận cỡ $m \times n$ bất kì. Khi đó

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T.$$

trong đó σ_i là các giá trị kì dị dương của ma trận $\mathbf{A}, \mathbf{u}_i$ là các vector kì dị trái và $\mathbf{v}_i$ là các vector kì dị phải của ma trận $\mathbf{A}$.

Thuật toán SVD

- Bước 1: Tính ma trận $\mathbf{A}^T\mathbf{A}$ và giải phương trình $\det\left(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}\right) = 0$ từ đó suy ra các giá trị kì dị của $\mathbf{A}$ là $\sigma_i = \sqrt{\lambda_i}, i = \overline{1,n}$ và $\mathbf{D} = \mathrm{diag}\left(\sigma_1, \sigma_2, \dots, \sigma_n\right)$.
- Bước 2: Với mỗi giá trị riêng λ_i, tìm vector riêng $\mathbf{v}_i$. Từ đó tìm được ma trận trực giao $\mathbf{V}$ cấp n chứa các vector kì dị phải của $\mathbf{A}$.
- Bước 3: Tìm ma trận trực giao $\mathbf{U}$ với

$$\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{A}\mathbf{v}_i.$$

3.1 Ý Tưởng

Tìm một phép xoay trục toạ độ để được một hệ trục toạ độ mới sao cho trong hệ mới này, thông tin của dữ liệu chủ yếu tập trung ở một vài thành phần. Phần còn lại chứa ít thông tin hơn có thể được lược bỏ.

Với $\mathbf{X} \in \mathbb{R}^{m \times n}$ với mỗi giá trị đã được chuẩn hóa sao cho mỗi hàng có giá trị trung bình là 0. Thì PCA của $\mathbf{X}$ với r thành phần là tìm ma trận trực giao $\mathbf{A} \in \mathbb{R}^{m \times r}$ và ma trận $\mathbf{B} \in \mathbb{R}^{n \times r}$ sao cho $\mathbf{X}$ có thể biểu diễn được dưới dạng ma trận.

$$\mathbf{X} = \mathbf{A}\mathbf{B}^T + \mathbf{E}.$$

trong đó $||\mathbf{E}||^2$ đạt giá trị nhỏ nhất.

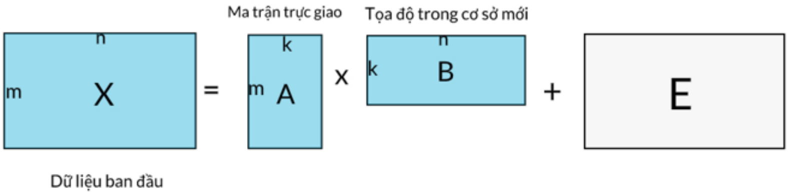

Figure: Ảnh minh họa PCA

Định lí

Nếu $(\mathbf{A}, \mathbf{B})$ là một nghiệm của mô hình PCA thì $(\mathbf{AQ}, \mathbf{BQ})$ cũng là một nghiệm của mô hình PCA, với $\mathbf{Q}$ là ma trận trực giao cấp r.
Lúc này, $\mathbf{Q}$ được gọi là phép quay trực giao.

Định nghĩa 3.2.2

Tập hợp các lớp tương đương $\overline{f}$ của $\mathbf{L_p}(X, \mathcal{A}, \mu)$ theo quan hệ "$\sim$" được xác định như trên được gọi là **không gian Lebesgue** $L_p(X, \mathcal{A}, \mu)$,

$$L_p(X, \mathcal{A}, \mu) = \{\overline{f} : f \in \mathbf{L_p}(X, \mathcal{A}, \mu)\}.$$

Kiểm tra ánh xạ $\|\cdot\|_p : L_p(X, \mathcal{A}, \mu) \to \mathbb{R}^+$ được xác định bởi

$$\|\overline{f}\|_p = \|f\|_p, \quad f \in \overline{f}$$

thỏa các Tiên đề N1, N2 của chuẩn. Tiên đề N3 có được từ bất đẳng thức Minkowski sau đây:

Bất đẳng thức Minkowski.

Cho $1 \leq p \leq \infty$ và $f, g \in L_p(X, \mathcal{A}, \mu)$. Khi đó $f + g \in L_p(X, \mathcal{A}, \mu)$ và

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Bất đẳng thức Minkowski.

Cho $1 \leq p \leq \infty$ và $f, g \in L_p(X, \mathcal{A}, \mu)$. Khi đó $f + g \in L_p(X, \mathcal{A}, \mu)$ và

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Bất đẳng thức Hölder.

Cho p, q là hai số thực không âm thoả $\frac{1}{p} + \frac{1}{q} = 1$ và $f \in L_p(X, \mathcal{A}, \mu)$, $g \in L_q(X, \mathcal{A}, \mu)$. Khi đó $fg \in L_1(X, \mathcal{A}, \mu)$ và

$$\int_X |fg| d\mu \leq \|f\|_p \|g\|_q.$$

Định lý 3.2.8

Cho $1 \leq p \leq \infty$. Khi đó, $(L_p(X, \mathcal{A}, \mu), \|\cdot\|_p)$ là không gian Banach.

Lấy $\{f_n\}_{n\in\mathbb{N}}$ là một dãy Cauchy trong $L_p(X, \mathcal{A}, \mu)$. Ta chỉ ra tồn tại một $f \in L_p(X, \mathcal{A}, \mu)$ và $f_n \to f$ trong $L_p(X, \mathcal{A}, \mu)$.

Định lý 3.2.7

Giả sử $f \in L_1(X, \mathcal{A}, \mu) \cap L_\infty(X, \mathcal{A}, \mu)$. Khi đó

(a) $f \in L_p(X, \mathcal{A}, \mu)$ với $1 < p < \infty$.

(b) $\lim_{p\to\infty} \|f\|_p = \|f\|_\infty$.

Định lý 3.2.9

Giả sử $\{f_n\}_{n\in\mathbb{N}}$ là một dãy hàm thuộc $L_p(X, \mathcal{A}, \mu)$, $1 \le p < \infty$, và $f_n \to f$ μ-hầu khắp nơi, trong đó $f \in L_p(X, \mathcal{A}, \mu)$. Khi đó

$$f_n \to f \iff \|f_n\|_p \to \|f\|_p.$$

Bổ đề 3.3.1.

Tập tất cả hàm đơn giản mà triệt tiêu bên ngoài một tập có độ đo hữu hạn là trù mật trong $L_p(X, \mathcal{A}, \mu)$, với $1 \leq p < \infty$.

• TH 1: $f \geq 0$. Khi đó, tồn tại dãy $\{s_n\}_{n \in \mathbb{N}}$ là dãy hàm đơn giản sao cho

$$0 \leq s_n \nearrow f.$$

Ta chỉ ra $s_n \to f$ trong $L_p(X, \mathcal{A}, \mu)$.

• TH 2: f tùy ý. Khi đó

$$f = f^+ - f^-,$$

với $f^+ = \max\{f, 0\}$ và $f^- = \max\{-f, 0\}$. Áp dụng TH 1.

Bổ đề 3.3.2.

Tập các hàm đơn giản trù mật trong $L_\infty(X, \mathcal{A}, \mu)$.

Sử dụng Định lý cấu trúc hàm đo được cho một hàm bị chặn hầu khắp nơi và bởi sự hội tụ trong $L_\infty(X, \mathcal{A}, \mu)$ chính là sự hội tụ đều hầu khắp nơi.

3.3 Các không gian trù mật trong $L_p(X, \mathcal{A}, \mu)$

Định lý 3.3.1.

Cho $f \in L_p(\mathbb{R}, \mathcal{L}, m)$ với $1 \leq p < \infty$. Khi đó, với mọi $\epsilon > 0$, tồn tại một hàm liên tục $g \in L_p(\mathbb{R}, \mathcal{L}, m)$ sao cho $\|f - g\|_p < \epsilon$.

Định lý 3.3.2.

Tập hợp tất cả các hàm bước nhảy trù mật trong $L_p(\mathbb{R}, \mathcal{L}, m)$ với $1 \leq p < \infty$.

Định lý 3.3.3.

Không gian $L_p(\mathbb{R}, \mathcal{L}, m)$ khả ly với $1 \leq p < \infty$.

$S = \{\sum_{k=1}^{n} b_k \chi_{J_k} : n \in \mathbb{N}, b_k \in \mathbb{Q}\}$, trong đó J_k,$1 \leq k \leq n$, là một khoảng hữu hạn với hai đầu mút hữu tỉ.

Bổ đề 3.4.1.

Cho $(X, \mathcal{A}, \mu)$ là một không gian độ đo hữu hạn. Giả sử $g \in L_1(X, \mathcal{A}, \mu)$ sao cho với bất kỳ $M > 0$ và mọi hàm đơn giản $s \in L_p(X, \mathcal{A}, \mu)$ xảy ra

$$\left| \int_X sg \, dm \right| \leq M \|s\|_p$$

với $1 \leq p < \infty$. Khi đó $g \in L_q(X, \mathcal{A}, \mu)$ và $\|g\|_q \leq M$.

Định lý Riesz

Cho $(X, \mathcal{A}, \mu)$ là một không gian độ đo σ-hữu hạn
• Với mỗi $g \in L_q(X, \mathcal{A}, \mu)$ thì xác định một phiếm hàm tuyến tính liên tục F trên $L_p(X, \mathcal{A}, \mu)$ được cho bởi

$$F(f) = \int_X fg\, d\mu,$$

và $\|F\| = \|g\|_q$.
• Với mỗi F là một phiếm hàm tuyến tính liên tục trên $L_p(X, \mathcal{A}, \mu)$. Khi đó tồn tại duy nhất hàm $g \in L_q(X, \mathcal{A}, \mu)$ sao cho

$$F(f) = \int_X fg\, d\mu, \quad f \in L_p(X, \mathcal{A}, \mu).$$

Hơn nữa $\|F\| = \|g\|_q$.

Ký hiệu $\mathfrak{BA}(X, \mathcal{A}, \mu)$ là tập hợp tất cả các độ đo có dấu cộng tính bị chặn và liên tục tuyệt đối với độ đo dương μ. Trên đó ta định nghĩa

$$\|\nu\|_{\text{var}} := |\nu|(X) = \sup\left\{\sum_{i=1}^{n} |\nu(A_i)| : (A_i)_{i=1}^{n} \subset \mathcal{A} \text{ và } X = \sqcup_{i=1}^{n} A_i\right\}.$$

Mệnh đề 3.4.2.

Cho $(X, \mathcal{A}, \mu)$ là một không gian độ đo. Khi đó $\|\cdot\|_{\text{var}}$ là một chuẩn trên $\mathfrak{BA}(X, \mathcal{A}, \mu)$.

Định nghĩa 3.4.2.

Cho $(X, \mathcal{A})$ là một không gian đo được và $\nu : \mathcal{A} \to \mathbb{R}$ là một độ đo có dấu cộng tính bị chặn. Với mọi f là một hàm đơn giản trên A, ta định nghĩa tích phân của f trên A ứng với ν là

$$\int_A f(x) d\nu(x) := \sum_{i=1}^{n} \alpha_i \nu(A_i),$$

ở đây $f = \sum_{i=1}^{n} \alpha_i \chi_{A_i}$ là biển diễn chính tắc của f trên A.

Định nghĩa 3.4.3.

Cho $(X, \mathcal{A})$ là một không gian đo được, $\nu : \mathcal{A} \to \mathbb{R}$ là một độ đo có dấu cộng tính bị chặn và $f : X \to \mathbb{R}$ là một hàm đo được và bị chặn trên X. Ta định nghĩa tích phân của f trên X ứng với ν cho bởi

$$\int_X f(x)d\nu(x) = \lim_{n\to\infty} \int_X f_n(x)d\nu(x),$$

trong đó $\{f_n\}_{n=1}^{\infty}$ là một dãy hàm đơn giản và hội tụ đều đến f trên X.

Bổ đề 3.4.3.

Cho $(X, \mathcal{A})$ là một không gian đo được, $\nu : \mathcal{A} \to \mathbb{R}$ là một độ đo có dấu cộng tính bị chặn. Giả sử $f : X \to \mathbb{R}$ là một hàm đo được và bị chặn trên X. Khi đó
a) Với mọi dãy hàm đơn giản $\{f_n\}_{n=1}^{\infty}$ thoả mãn $f_n(x) \rightrightarrows f(x)$ trên X thì tồn tại giới hạn

$$\lim_{n\to\infty} \int_X f_n(x) d\nu(x).$$

b) Nếu $\{f_n\}_{n=1}^{\infty}$ và $\{g_n\}_{n=1}^{\infty}$ là hai dãy hàm đơn giản thoả mãn $f_n(x) \rightrightarrows f(x)$ và $g_n(x) \rightrightarrows f(x)$ trên X thì

$$\lim_{n\to\infty} \int_X f_n(x) d\nu(x) = \lim_{n\to\infty} \int_X g_n(x) d\nu(x).$$

Định nghĩa 3.4.4.

Cho $(X, \mathcal{A}, \mu)$ là một không gian độ đo và $\nu \in \mathfrak{BA}(X, \mathcal{A}, \mu)$. Với mọi $f \in L_\infty(X, \mathcal{A}, \mu)$ ta định nghĩa tích phân của f trên X ứng với ν cho bởi

$$\int_X f(x)d\nu(x) = \lim_{n\to\infty} \int_X f_n(x)d\nu(x),$$

với $\{f_n\}_{n=1}^{\infty}$ là một dãy hàm đơn giản và hội tụ đều đến f trên $X \setminus Y$ trong đó Y là một tập có độ đo không.

Mệnh đề 3.4.4.

Nếu $f = g$ μ-hầu khắp nơi, thì

$$\int_X f(x)d\nu(x) = \int_X g(x)d\nu(x).$$

Mệnh đề 3.4.5.

Toán tử tích phân trong Định nghĩa 3.4.4 là toán tử tuyến tính.

Mệnh đề 3.4.6.

Cho $(X, \mathcal{A}, \mu)$ là một không gian độ đo và $\nu \in \mathfrak{BA}(X, \mathcal{A}, \mu)$. Khi đó

$$\left| \int_X f(x) d\nu(x) \right| \le \|\nu\|_{\text{var}} \|f\|_\infty$$

với mọi $f \in L_\infty(X, \mathcal{A}, \mu)$.

Định lý biểu diễn Kantorovich.

Cho $(X, \mathcal{A}, \mu)$ là một không gian độ đo. Khi đó
a) Với mỗi $\nu \in \mathfrak{BA}(X, \mathcal{A}, \mu)$, thì phiếm hàm $F : L_\infty(X, \mathcal{A}, \mu) \to \mathbb{R}$ xác định bởi

$$F(f) = \int_X f(x) d\nu(x), \qquad \forall f \in L_\infty(X, \mathcal{A}, \mu)$$

là một phiếm hàm tuyến tính liên tục trên $L_\infty(X, \mathcal{A}, \mu)$.
b) Với mỗi F là phiếm hàm tuyến tính liên tục trên $L_\infty(X, \mathcal{A}, \mu)$ tồn tại duy nhất phần tử $\nu \in \mathfrak{BA}(X, \mathcal{A}, \mu)$ sao cho

$$F(f) = \int_X f(x) d\nu(x), \qquad \forall f \in L_\infty(X, \mathcal{A}, \mu).$$

Hơn nữa $\|\nu\|_{\text{var}} = \|F\|$.

Định lý 3.5.1.

Cho $(X, \mathcal{A}, \mu)$ là một không gian độ đo hữu hạn và $1 < p < \infty$. Đặt $\{f_n\}_{n\in\mathbb{N}}$ là một dãy hàm trong $L_p(X, \mathcal{A}, \mu)$ sao cho $f_n \to f$ μ-hầu khắp nơi. Khi đó $f_n \rightharpoonup f$ khi và chỉ khi $\{\|f_n\|_p\}_{n\in\mathbb{N}}$ bị chặn.

Định lý 3.5.1.

Cho $(X, \mathcal{A}, \mu)$ là một không gian độ đo hữu hạn và $1 < p < \infty$. Đặt $\{f_n\}_{n\in\mathbb{N}}$ là một dãy hàm trong $L_p(X, \mathcal{A}, \mu)$ sao cho $f_n \to f$ μ-hầu khắp nơi. Khi đó $f_n \rightharpoonup f$ khi và chỉ khi $\{\|f_n\|_p\}_{n\in\mathbb{N}}$ bị chặn.

Định lý 3.5.2.

Giả sử $\{f_n\}_{n\in\mathbb{N}} \subset L_p(\mathbb{R})$ với $M = \sup_n \|f_n\|_p < \infty$. Khi đó tồn tại $f \in L_p(\mathbb{R})$ sao cho $\|f\|_p \leq M$ và một dãy con $\{f_{n_k}\}_{k\in\mathbb{N}} \subset \{f_n\}_{n\in\mathbb{N}}$ sao cho

$$f_{n_k} \rightharpoonup f.$$

Tài liệu tham khảo

- L. Đ. Kỳ, *Bài giảng về Lý thuyết độ đo và tích phân*, Quy Nhơn, 2018.
- L. Đ. Kỳ, *Bài giảng về Giải tích thực*, Quy Nhơn, 2018.
- T.T. Quang, *Cơ sở lý thuyết Giải tích hàm*, Quy nhơn, 2013.
- H. Brezis, *Functional analysis*, Sobolev spaces and partial differential equations. Universitext. Springer, New York, 2011.
- R. E. Castillo and H. Rafeiro, *An introductory course in Lebesgue spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, [Cham], 2016.
- H. L. Royden, *Real analysis*. Fourth edition. Pearson Education and China Machine Press, China, 2010.

KẾT THÚC BÁO CÁO

TRÂN TRỌNG CẢM ƠN!