Análisis de Componentes Principales

Lección 27

Dr. Pablo Alvarado Moya

CE3102 Análisis Numérico para Ingeniería
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

I Semestre 2018

Contenido

Introducción

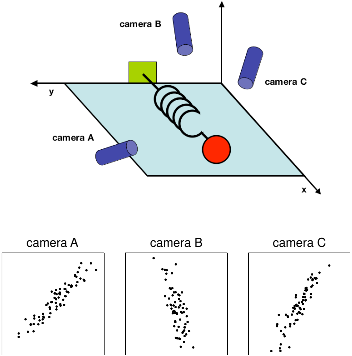
- El Análisis de Componentes Principales (ACP, o PCA) es una herramienta estándar del análisis de datos.
- Sus aplicaciones se encuentran desde neurociencias hasta gráficos por computador.
- Su tarea principal e reducir el número de dimensiones de un conjunto de datos, con el afán de revelar estructuras simples usualmente ocultas en esos datos.
- Referencia: Shlens, J. A Tutorial on PCA.

Experimento de partida (1)

- Supóngase que realizamos un experimento en donde se miden varias cantidades para analizar un fenómeno (p. ej. espectros, tensiones eléctricas, velocidades, etc.)
- Los datos medidos pueden aparecer difusos, redundantes, poco claros, y es en general un obstáculo en ciencias con validaciones empíricas.

Experimento de partida (2)

- Supóngase que se analiza el montaje para estudiar el movimiento de un resorte ideal.



Experimento de partida (3)

Una bola de masa m está sujeta a un resorte sin masa y sin fricción. La bola se libera a una pequeña distancia alejada del equilibrio. La bola oscila entonces en el eje x a una determinada frecuencia.

- Supóngase que se ignora cuáles ejes son importantes para la medición, así que se toman medidas en el espacio tridimensional utilizando tres cámaras de alta velocidad que toman 120 cuadros por segundo.
- Las cámaras se colocan en tres posiciones y orientaciones arbitrarias.
- La pregunta es cómo obtener de todas las mediciones la ecuación que ponen en evidencia la única dependencia de x.
- En el mundo real, las mediciones tomadas están además contaminadas con ruido, y por efectos de condiciones no ideales.

Cambio de base (1)

- La meta del ACP es identificar la base con "mayor" sentido para reexpresar los datos.
- Se espera además que en esta nueva base sea fácil filtrar los efectos del ruido y encontrar estructuras ocultas en los datos (por ejemplo, descubrir de las mediciones que $\underline{\mathbf{x}}$ es la dimensión de interés.

Una base canónica (1)

- Se tratará a cada dato tomado en el experimento como un elemento del conjunto de datos. A este dato se le denomina **muestra**.
- El dato es un vector conformado por todas las mediciones tomadas (tensión eléctrica, posición, etc.)
- En el ejemplo, la cámara A reporta una posición (x_A, y_A), así que una muestra en el ejemplo está dada por:

$$\underline{\mathbf{x}} = \begin{bmatrix} x_A & y_A & x_B & y_B & x_C & y_C \end{bmatrix}^T$$

- Si se grabaran estas posiciones durante 10 minutos a 120 fps se tendrían entonces 72000 muestras.

Una base canónica (2)

- Cada muestra $\underline{\mathbf{x}}$ es un vector en un espacio m-dimensional, con m el número de mediciones tomados.
- Dicho espacio es engendrado por una base ortogonal, que por convención está formada por los vectores $\underline{\mathbf{b}}_1 = [1, 0, \ldots, 0]^T$, $\underline{\mathbf{b}}_2 = [0, 1, 0, \ldots, 0]^T$, ..., $\underline{\mathbf{b}}_n = [0, \ldots, 0, 1]^T$
- La matriz de proyección a esta base es entonces

$$\mathbf{B} = \begin{bmatrix} \underline{\mathbf{b}}_1^T \\ \underline{\mathbf{b}}_2^T \\ \vdots \\ \underline{\mathbf{b}}_n^T \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}$$

- Esta base “ingenua”, o formalmente, base **canónica**, mantiene los datos exactamente como se han medido.

Cambio de base (1)

- La pregunta es ahora ¿Existe alguna combinación **lineal** de la base original que exprese "mejor" el conjunto de datos?
- Sea $\mathbf{X}$ el conjunto original de datos, donde cada columna es una muestra.
- En el ejemplo anterior $\mathbf{X}$ es una matriz 6×72000.
- Sea $\mathbf{Y}$ otra matriz resultante de la proyección o transformación lineal $\mathbf{P}$:

$$\mathbf{Y} = \mathbf{P}\mathbf{X}$$

- Sea $\underline{\mathbf{p}}_i^T$ una fila de $\mathbf{P}$
- Sea $\underline{\mathbf{x}}_j$ una columna de $\mathbf{X}$ (la j-ésima muestra)

Cambio de base (2)

- Sea $\underline{\mathbf{y}}_j$ una columna de $\mathbf{Y}$ (la j-ésima muestra)
- $\mathbf{P}$ es una matriz que transforma $\mathbf{X}$ en $\mathbf{Y}$
- Geométricamente, $\mathbf{P}$ es una rotación y estiramiento que transforma $\mathbf{X}$ en $\mathbf{Y}$
- Las filas de $\mathbf{P}$ son la nueva base del espacio vectorial, lo que se observa con

$$\mathbf{PX} = \begin{bmatrix} \underline{\mathbf{p}}_1^T \\ \vdots \\ \underline{\mathbf{p}}_m^T \end{bmatrix} \begin{bmatrix} \underline{\mathbf{x}}_1 & \cdots & \underline{\mathbf{x}}_n \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} \underline{\mathbf{p}}_1^T \underline{\mathbf{x}}_1 & \cdots & \underline{\mathbf{p}}_1^T \underline{\mathbf{x}}_n \\ \vdots & \ddots & \vdots \\ \underline{\mathbf{p}}_m^T \underline{\mathbf{x}}_1 & \cdots & \underline{\mathbf{p}}_m^T \underline{\mathbf{x}}_n \end{bmatrix}$$

Cambio de base (3)

es decir, la i-ésima columna de $\mathbf{Y}$

$$\underline{\mathbf{y}}_i = \begin{bmatrix} \underline{\mathbf{p}}_1^T \underline{\mathbf{x}}_i \\ \underline{\mathbf{p}}_2^T \underline{\mathbf{x}}_i \\ \vdots \\ \underline{\mathbf{p}}_m^T \underline{\mathbf{x}}_i \end{bmatrix}$$

se conforma por la proyección del i-ésimo vector original $\underline{\mathbf{x}}_i$ sobre cada uno de los vectores $\underline{\mathbf{p}}_j$

- La preguntas que quedan por responder son ¿cómo deben ser elegidos los vectores de la nueva base? ¿qué es bueno rescatar de $\mathbf{X}$?, o similar ¿qué debería contener $\mathbf{Y}$?

Varianza y valor medio (1)

- El valor medio o promedio $\underline{\boldsymbol{\mu}}$ del conjunto de datos se estima como:

$$\underline{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \underline{\mathbf{x}}_i$$

- La varianza σ^2 se estima como el promedio de desviaciones de los datos con respecto a su valor medio

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \|\underline{\mathbf{x}}_i - \underline{\boldsymbol{\mu}}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} (\underline{\mathbf{x}}_i - \underline{\boldsymbol{\mu}})^T(\underline{\mathbf{x}}_i - \underline{\boldsymbol{\mu}})$$

- El cálculo de estas magnitudes se puede realizar con **un** solo recorrido por el conjunto de datos.
- **¿Cómo?**

Varianza y valor medio (2)

- Para demostrar qué debe calcularse para encontrar sigma con un solo recorrido de datos.

$$
\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(\underline{\mathbf{x}}_i - \underline{\boldsymbol{\mu}})^T(\underline{\mathbf{x}}_i - \underline{\boldsymbol{\mu}}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\underline{\mathbf{x}}_i^T\underline{\mathbf{x}}_i - \underline{\mathbf{x}}_i^T\underline{\boldsymbol{\mu}} - \underline{\boldsymbol{\mu}}^T\underline{\mathbf{x}}_i + \underline{\boldsymbol{\mu}}^T\underline{\boldsymbol{\mu}}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\|\underline{\mathbf{x}}_i\|^2 - 2\underline{\boldsymbol{\mu}}^T\underline{\mathbf{x}}_i + \|\underline{\boldsymbol{\mu}}\|^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\|\underline{\mathbf{x}}_i\|^2 - 2\underline{\boldsymbol{\mu}}^T\frac{1}{n}\sum_{i=1}^{n}\underline{\mathbf{x}}_i^T + \|\underline{\boldsymbol{\mu}}\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\|\underline{\mathbf{x}}_i\|^2 - \|\underline{\boldsymbol{\mu}}\|^2
\end{aligned}
$$

Ruido y rotación (1)

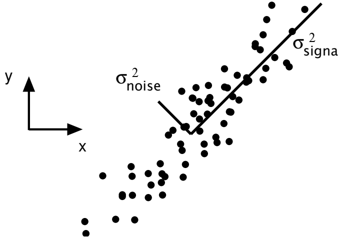
- El ruido de medición en cualquier conjunto de datos debe ser lo suficientemente bajo para poder rescatar la información en ellos.
- La calidad de los datos (o señal) se mide con la tasa de señal a ruido (SNR), descrita usualmente a través de una tasa de varianzas:

$$SNR = \frac{\sigma^2_{\text{señal}}}{\sigma^2_{\text{ruido}}}$$

  es decir, el ruido se mide respecto a la señal, usualmente "potencia" o energía entre ambas.
- SNR $\gg 1$ indica mediciones de alta precisión
- Baja SNR indica datos ruidosos.

Ruido y rotación (2)

- Volviendo al ejemplo del resorte.



- Se supone que el resorte mueve la bola en una línea recta, así que las tres cámaras deberían observar un movimiento en línea recta.
- Toda desviación de la línea recta es ruido.

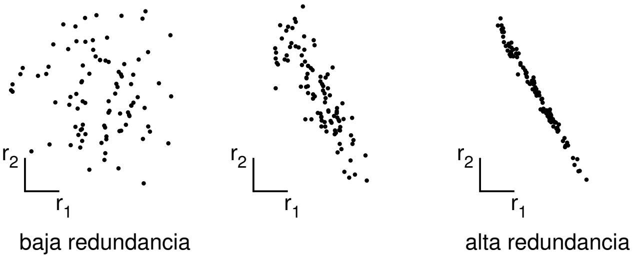
Ruido y rotación (3)

- La varianza asociada a la señal y al ruido se demarcan en el diagrama.
- Si se asume que la señal es suficientemente fuerte, entonces la dirección de la **mayor varianza** coincide con la señal.
- Se busca entonces una rotación de los vectores de la base canónica para alinearlos con los ejes de mayor varianza.

Redundancia (1)

- Otro factor a considerar es la **redundancia**.
- En el ejemplo de las cámaras es notorio, pues cada cámara registra la misma información.
- Obsérvese que incluso en una misma imagen, si sube x, también sube y.
- Existen diversas dependencias entre dos variables r_1 y r_2:

Redundancia (2)



- En imagen anterior con una dimensión es suficiente para describir los datos altamente redundantes (se puede reducir la dimensión), pero dos dimensiones son necesarias si los datos en r_1 y r_2 no redundan.

Matriz de covarianza (1)

- Considérense dos conjuntos de mediciones con media cero

$$A = \{a_1, a_2, \ldots, a_n\} \qquad B = \{b_1, b_2, \ldots, b_n\}$$

- Las varianzas de los datos A y B por separado son

$$\sigma_A^2 = \frac{1}{n}\sum_i a_i^2, \qquad \sigma_B^2 = \frac{1}{n}\sum_i b_i^2$$

Matriz de covarianza (2)

- La **covarianza** entre A y B se generaliza como

$$\sigma_{AB}^2 = \frac{1}{n} \sum_i a_i b_i$$

- La covarianza mide el grado de dependencia lineal entre dos variables:
  - un valor positivo grande indica datos correlacionados positivamente
  - un valor negativo grande indica datos correlacionados negativamente
  - es cero si A y B no están correlacionados (no hay redundancia)
  - $\sigma_{AB}^2 = \sigma_A^2$ si $A = B$
- El valor absoluto de la covarianza indica el grado de redundancia

Matriz de covarianza (3)

- Expresando A y B como vectores:

$$A \rightarrow \underline{\mathbf{a}}^T = [a_1, a_2, \ldots, a_n]$$
$$B \rightarrow \underline{\mathbf{b}}^T = [b_1, b_2, \ldots, b_n]$$

se replantea la covarianza como

$$\sigma^2_{\underline{\mathbf{ab}}} = \frac{1}{n}\underline{\mathbf{a}}^T\underline{\mathbf{b}}$$

Matriz de covarianza (4)

- Lo anterior se puede generalizar de dos a cualquier número de vectores $\underline{\mathbf{x}}_i$: Dada la matriz $\mathbf{X}$ de dimensiones $m \times n$

$$\mathbf{X} = \begin{bmatrix} \underline{\mathbf{x}}_1^T \\ \underline{\mathbf{x}}_2^T \\ \vdots \\ \underline{\mathbf{x}}_m^T \end{bmatrix}$$

  - Las filas de $\mathbf{X}$ corresponden a todas las mediciones de un tipo particular (por ejemplo, eje x de la cámara A).
  - Cada columna de $\mathbf{X}$ corresponde a un conjunto de mediciones de una única muestra.

Matriz de covarianza (5)

- La matriz de covarianza $\Sigma_{\mathbf{X}}$ se define entonces como

$$\Sigma_{\mathbf{X}} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$$

- El elemento (i, j) de la matriz $\Sigma_{\mathbf{X}}$ es el producto punto entre el i-ésimo tipo de medición y el j-ésimo tipo.
- Esta matriz es
  - Cuadrada $(m \times m)$ y simétrica
  - La diagonal está formada por las varianzas de cada tipo de medición
  - Los términos fuera de la diagonal son las covarianzas entre tipos de medición.
- Los valores de covarianzas reflejan niveles de ruido y redundancia en las mediciones.

Matriz de covarianza (6)

- En la diagonal, valores de gran magnitud indican estructuras relevantes
- Fuera de la diagonal, valores grandes en magnitud indican alta redundancia.

Diagonalización de la matriz de covarianza (1)

- Es de interés minimizar la redundancia y maximizar la señal de los datos de salida $\mathbf{Y}$, y por lo tanto
  - Todos los datos fuera de la diagonal de $\Sigma_{\mathbf{Y}}$ deben ser cero, por lo que $\Sigma_{\mathbf{Y}}$ es diagonal
  - Cada dimensión sucesiva de $\mathbf{Y}$ deberá estar ordenada de acuerdo a la varianza.
- La matriz $\Sigma_{\mathbf{X}}$ debe ser entonces **diagonalizada**
- El ACP asume que los vectores de la nueva base son ortonormales, y éstos constituyen los **componentes principales**.
- La varianza en cada componente indica qué tan importante es el componente.

Resumen de supuestos

- Se ha supuesto que el problema es **lineal**
  Un cambio de base expresado como producto de matrices puede solucionar el problema.
- Varianzas grandes indican estructura importante.
  Esto supone que se tiene alto SNR, para que las varianzas grandes estén asociadas con la señal.
- Componentes principales son ortogonales.
  Esto permite encontrar una solución utilizando álgebra lineal.

Derivación de diagonalización (1)

- El problema que se plantea es: dado el conjunto de datos $\mathbf{X}$ (matriz $m \times n$, cada columna representa una muestra de m dimensiones, se cuenta con n muestras), encuéntrese una matriz ortogonal $\mathbf{P}$ que transforma los datos a $\mathbf{Y} = \mathbf{P}\mathbf{X}$, de tal modo que

$$\Sigma_{\mathbf{Y}} = \frac{1}{n}\mathbf{Y}\mathbf{Y}^T$$

  sea una matriz diagonal.
- Las filas de $\mathbf{P}$ son los componentes principales de $\mathbf{X}$

Derivación de diagonalización (2)

- Se parte de $\Sigma_{\mathbf{Y}}$

$$\begin{aligned}
\Sigma_{\mathbf{Y}} &= \frac{1}{n}\mathbf{Y}\mathbf{Y}^T \\
&= \frac{1}{n}(\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^T \\
&= \frac{1}{n}\mathbf{P}\mathbf{X}\mathbf{X}^T\mathbf{P}^T \\
&= \mathbf{P}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T\right)\mathbf{P}^T \\
&= \mathbf{P}\Sigma_{\mathbf{X}}\mathbf{P}^T
\end{aligned}$$

que relaciona las matrices de covarianza de datos de entrada y salida

Derivación de diagonalización (3)

- Ya se demostró lecciones atrás que lo anterior se cumple si las columnas de $\mathbf{P}^T$ son los eigenvectores de $\mathbf{X}$, y los elementos en la matriz diagonal $\Sigma_\mathbf{Y}$ los eigenvalores.
- En otras palabras, las filas de $\mathbf{P}$ son los eigenvectores, que se ordenan de acuerdo a sus correspondientes eigenvalores de forma descendiente.
- Los componentes principales de $\mathbf{X}$ están dados entonces por los eigenvectores de su matriz de covarianza $\Sigma_\mathbf{X} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$
- El i-ésimo valor de la diagonal de $\Sigma_\mathbf{Y}$ es la varianza de $\mathbf{X}$ a lo largo del componente principal $\mathbf{p}_i$.

Pasos del ACP

1. Substraer la media de todos los datos
2. Calcular la matriz de covarianza $\Sigma_{\mathbf{X}}$
3. Calcular los eigenvectores y eigenvalores de $\Sigma_{\mathbf{X}}$
4. Construir la matriz de proyección $\mathbf{P}$ compuesta por el número de componentes principales deseados (seleccionados de acuerdo a los eigenvalores)
5. Proyectar los datos con $\mathbf{Y} = \mathbf{PX}$

Reconstrucción

- Para aproximar/reducir la dimensión usamos:

$$\underline{\mathbf{y}} = \mathbf{P}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})$$

- Para reconstruir el dato original $\underline{\mathbf{x}}$ a partir de la representación $\underline{\mathbf{y}}$ hacemos:

$$\underline{\mathbf{x}} \approx \underline{\tilde{\mathbf{x}}} = \mathbf{P}^T \underline{\mathbf{y}} + \underline{\boldsymbol{\mu}}$$

Ejemplos

- Eigenfaces
  http://www.youtube.com/watch?v=5nBL_u4MF0k
- AAM http://www.youtube.com/watch?v=I3YsqHCQB4k

Resumen

FIN

Este documento ha sido elaborado con software libre incluyendo LATEX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, LTI-Lib-2, GNU-Make y Subversion en GNU/Linux

Este trabajo se encuentra bajo una Licencia Creative Commons Atribución-NoComercial-LicenciarIgual 3.0 Unported. Para ver una copia de esta Licencia, visite http://creativecommons.org/licenses/by-nc-sa/3.0/ o envíe una carta a Creative Commons, 444 Castro Street, Suite 900, Mountain View, California, 94041, USA.

© 2005-2018 Pablo Alvarado-Moya Área de Ingeniería en Computadores Instituto Tecnológico de Costa Rica