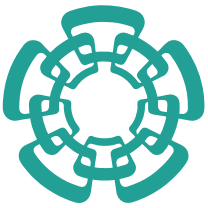

Centro de Investigación y de Estudios Avanzados del IPN
Laboratorio de Tecnologías de Información

PCA
Principal Component Analysis

Rafael Pérez Torres
Tópicos selectos en reconocimiento de patrones
Profesor: Dr. Wilfrido Gómez

3 julio 2015

Resumen

El presente documento presenta una introducción a las técnicas de reducción de dimensionalidad, describiendo tópicos como la maldición de la dimensionalidad así como una clasificación general de las técnicas de reducción. Específicamente, se describe el Análisis de Componentes Principales (PCA) como una técnica no paramétrica para la reducción de dimensionalidad, presentando su fundamento matemático y el método de descomposición de eigenvalores como mecanismo para selección de los componentes principales.

1. Introducción

La maldición de la dimensionalidad La maldición de la dimensionalidad es un término acuñado por Bellman en 1961, y se refiere al hecho de que el tamaño de la muestra para estimar una función de varias *variables* con un grado específico de precisión aumenta exponencialmente con el número de variables.

El fenómeno del espacio vacío Acuñado por Scott y Thompson, el fenómeno del espacio vacío es responsable de la maldición de la dimensionalidad, indicando que los espacios de altas dimensiones son inherentemente dispersos (*sparse*).

Dimensión intrínseca La dimensión intrínseca de un fenómeno es el número de variables independientes que explican o describen de forma satisfactoria a dicho fenómeno. A partir de ella se puede realizar la reducción de dimensionalidad sin caer en casos extremos.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_3 \end{bmatrix}\right)$$

(a) Generación de características

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \rightarrow \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iM} \end{bmatrix}$$

(b) Selección de características

Figura 1: Reducción de dimensionalidad

Reducción de la dimensionalidad Dada una cantidad de características, ¿Cómo pueden seleccionarse las más importantes de forma que se reduzca su cantidad y al mismo tiempo se conserve la mayor cantidad de información discriminatoria posible?

¿Por qué es posible reducir la dimensionalidad? A menudo, la representación de los datos será posible por distintos motivos:

- Muchas de las variables son menores que el tamaño del ruido por lo que son irrelevantes.
- Muchas de las variables están correlacionadas con otras.

Tipos de reducción de dimensionalidad Existen dos mecanismos principales para reducir la cantidad de variables en un sistema (Figura 1):

- **Extracción-Generación de características**: Dado un espacio de características $\mathbf{x}_i \in \mathbb{R}^M$, encontrar un mapeo $\mathbf{y} = f(x) : \mathbb{R}^M \rightarrow \mathbb{R}^m$, con $m < M$, tal que el vector transformado $\mathbf{y}_i \in \mathbb{R}^m$ preserve la información en $\mathbb{R}^M$.
- **Selección de características**: Dado un espacio de características $\mathbf{x}_i = \{x_j | j = 1, \ldots, M\}$ encontrar un subconjunto $\mathbf{y}_i = \{x_{i1}, \ldots, x_{im}\}$, con $m < M$, tal que se maximice el desempeño de la clasificación.

2. Principal Component Analysis PCA

El Análisis de Componentes Principales es una herramienta estándar en el análisis de datos moderno debido a que es no paramétrico, es simple, y existen algoritmos relativamente eficientes para su cálculo.

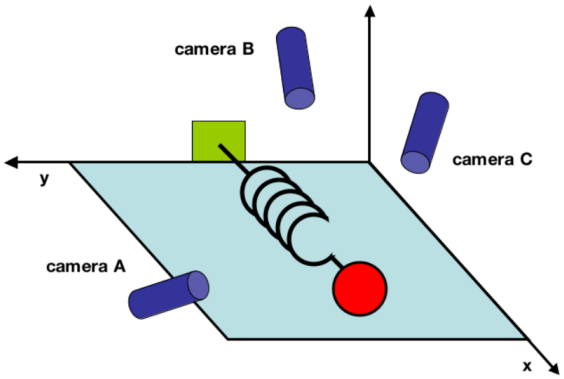

Figura 2: Sistema para estimación del comportamiento de un resorte

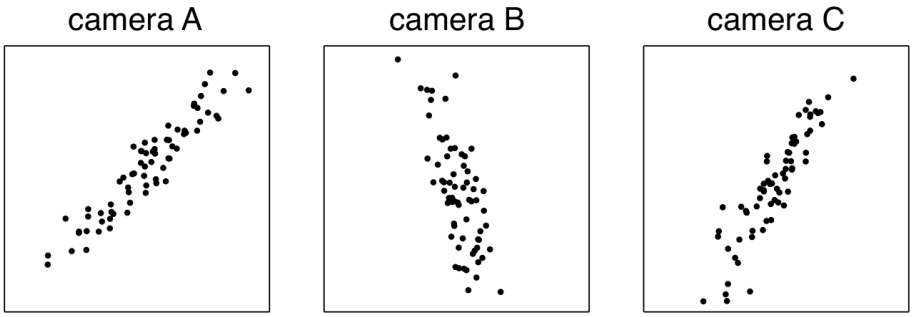

Figura 3: Datos obtenidos por las cámaras

En términos simples, la técnica PCA busca identificar las bases más significativas para reexpresar un conjunto de datos que filtre el ruido y revele estructuras ocultas.

2.1. Caso de uso básico

Considérese que se desea estudiar el movimiento de un resorte ideal. Asumiendo que se ignoran conocimientos de física, se prepara un dispositivo experimental que consiste en el resorte y tres cámaras de video colocadas en posiciones arbitrarias ($\overrightarrow{a}$, $\overrightarrow{b}$, $\overrightarrow{c}$), como se muestra en la Figura 2.

Intuitivamente sabemos que el movimiento se concentra únicamente en el eje X, pero ¿Cómo pasar de lo detectado por las cámaras en la Figura 3 hacia esto?

En el mundo real, típicamente se trabaja *a ciegas* y no es posible conocer con anticipación la mejor posición en la que las *cámaras* deban ser colocadas.

$$v = k_1 v_1 + k_2 v_2 + \ldots + k_n v_n = \sum_{i=1}^{n} k_i v_i$$

Figura 4: Ejemplo de combinación lineal

$$\mathbf{PX} = \underbrace{\begin{bmatrix} p_1 \\ \vdots \\ p_m \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}}_{\text{(a)}} \quad Y = \underbrace{\begin{bmatrix} p_1 x_1 & \ldots & p_1 x_n \\ \vdots & \ddots & \vdots \\ p_m x_1 & \cdots & p_m x_n \end{bmatrix}}_{\text{(b)}}$$

Figura 5: Nuevos vectores base

2.2. Fundamentos matemáticos del PCA

PCA asume que los datos puedes ser representados como una combinación lineal de sus vectores básicos (Figura 4).

Sea $\mathbf{X}$ una matriz $m \times n$ el conjunto original de datos, $\mathbf{Y}$ otra matriz $m \times n$ que almacenará la nueva representación de los datos realizada a través de una transformación lineal con la matriz $\mathbf{P}$. Entonces se tiene:

$$\mathbf{PX} = \mathbf{Y} \tag{1}$$

A partir de 1 es posible interpretar que:

- Geométricamente, P es una rotación y estrechamiento que transforma a $\mathbf{X}$ en $\mathbf{Y}$
- Las filas de $\mathbf{P}$, $\{p_1, \ldots, p_m\}$ son el nuevo conjunto de vectores base para expresar a los elementos de $\mathbf{X}$, como se muestra en la Figura 5.
- El j-ésimo coeficiente de y_i es una proyección sobre la j-ésima fila de $\mathbf{P}$

De lo anterior se desprende que los vectores fila $\{p_1, \ldots, p_m\}$ en $\mathbf{P}$ serán los componentes principales de $\mathbf{X}$. Para encontrar la mejor forma para reexpresar $\mathbf{X}$ así como una buena elección para $\mathbf{P}$ es necesario identificar las características que nos gustaría que $\mathbf{Y}$ exhiba.

Elementos para describir los datos

Ruido y rotación El nivel del ruido en cualquier conjunto de datos debería ser bajo. No existe una escala absoluta para su medición pero sí para obtener una referencia cuantitativa respecto a la magnitud de la *señal* descrita por los datos. Una métrica común es el radio señal a ruido (SNR), o un radio de varianzas σ^2,

$$SNR = \frac{\sigma^2_{\text{signal}}}{\sigma^2_{\text{noise}}}$$

Un valor de SNR alto (> 1) indica una medición de alta precisión, mientras que un valor bajo indica datos muy ruidosos. Esta relación es mostrada en la

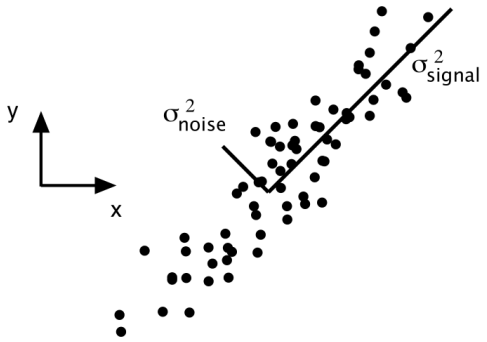

Figura 6: Relación entre SNR y varianza de los datos

Figura 7: Niveles de redundancia

Figura 6, donde puede apreciarse que la dirección de mayor varianza coincide con el eje mayor de la nube de puntos; así se infiere que la dinámica de interés ocurre en la dirección con la mayor varianza y presumiblemente el más alto valor de SNR.

Redundancia La redundancia entre características ocurre cuando a partir del valor de un atributo $r1$ es posible calcular el valor de un atributo $r2$. La Figura 7 muestra diferentes niveles de redundancia entre dos atributos. Idealmente, se busca considerar la menor cantidad de variables, por lo que aquellas redundantes son discriminadas.

Matriz de covarianza La covarianza permite conocer el grado de relación lineal entre dos variables. Un gran valor positivo indica correlación positiva, mientras que un gran valor negativo denota correlación negativa. Por consiguiente, el valor absoluto de la covarianza mide el grado de redundancia. Si cada fila de $\mathbf{X}$ representa todas las mediciones de un tipo, cada columna corresponde al

conjunto de medidas tomadas en una lectura en particular:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

entonces la matriz de covarianza $mathbf{C}_X$ puede ser expresada como:

$$C_X \equiv \frac{1}{n}\mathbf{X}\mathbf{X}^T$$

donde el ij-ésimo elemento de $\mathbf{C_X}$ corresponde al producto punto entre el vector i-ésimo y el vector j-ésimo de $\mathbf{X}$. La matriz $\mathbf{C_X}$ es cuadrada de tamaño $m \times m$ además de simétrica y permite conocer los valores de ruido y redundancia de los datos sabiendo que:

- Su diagonal incluye la varianza de cada característica, valores grandes indican importancia estructural.
- Los elementos no diagonales definen la covarianza entre cada par de atributos, valores grandes indican alta redundancia.

De esta manera, en la nueva matriz de covarianza $\mathbf{C_Y}$ se desea contener los datos reestructurados intentando a) minimzar redundancia (medida mediante la covarianza) y b) maximizar la señal (medida mediante la varianza). Por ello, la matriz $\mathbf{C_Y}$ debería tener la siguiente estructura:

- Los elementos no diagonales deberían ser 0, convirtiéndola en una matriz diagonal que es *de-correlacionada*.
- Cada dimensión en $\mathbf{Y}$ debería ser ordenada de acuerdo a su varianza.

El elemento restante de la ecuación inicial 1 es $\mathbf{P}$ quien se comporta como una rotación para alinear los datos con el eje de máxima varianza. La idea general de su cálculo es mostrada en el Algoritmo 1.

Algoritmo 1 Algoritmo para encontrar el valor de $\mathbf{P}$

1: Seleccionar una dirección en el espacio m-dimensional en el que la varianza de $\mathbf{X}$ es maximizada y guardar dicho vector como $\mathbf{p_1}$.
2: Encontrar otra dirección donde la varianza sea maximizada, restringiendo la búsqueda a las direcciones ortogonales a todas las direcciones previamente seleccionadas. Almacenar este nuevo vector como $\mathbf{p_i}$.
3: Repetir el procedimiento hasta que m vectores hayan sido seleccionados.

Al ordenar el conjunto resultante en $\mathbf{p}$ se obtienen los componentes principales.

3. Solución de PCA a través de descomposición de eigenvectores

El objetivo de la solución es encontrar una matriz ortonormal $\mathbf{P}$ en $\mathbf{Y} = \mathbf{PX}$ tal que $\mathbf{C_Y} \equiv \frac{1}{n}\mathbf{YY}^T$. Las filas de $\mathbf{P}$ son los componentes principales de $\mathbf{X}$.

Expresando $\mathbf{C_Y}$ en términos de la variable desconocida $\mathbf{P}$:

$$\begin{aligned}
\mathbf{C_Y} &= \frac{1}{n}\mathbf{YY}^T && (2)\\
&= \frac{1}{n}(\mathbf{PX})(\mathbf{PX})^T && (3)\\
&= \frac{1}{n}\mathbf{PXX}^T\mathbf{P}^T && (4)\\
&= P(\frac{1}{n}\mathbf{XX}^T)\mathbf{P}^T && (5)\\
\mathbf{C_Y} &= \mathbf{PC_X}\mathbf{P}^T && (6)
\end{aligned}$$

Cualquier matriz simétrica $\mathbf{A}$ es diagonalizada por una matriz ortogonal[1] de sus eigenvectores, esto es $\mathbf{A} = \mathbf{EDE}^T$. Al seleccionar que la matriz $\mathbf{P}$ sea una matriz donde cada renglón $\mathbf{p_i}$ es un eigenvector de $\frac{1}{n}\mathbf{XX}^T$, se logra que $\mathbf{P} \equiv \mathbf{E^T}$. Adicionalmente se debe considerar que $\mathbf{P^{-1}} = \mathbf{P^T}$ Utilizando estos elementos, es posible terminar de evaluar $\mathbf{C_Y}$ como:

$$\begin{aligned}
\mathbf{C_Y} &= \mathbf{PC_X}\mathbf{P}^T && (7)\\
&= \mathbf{P}(\mathbf{E^T}\mathbf{DE})\mathbf{P^T} && (8)\\
&= \mathbf{P}(\mathbf{P^T}\mathbf{DP})\mathbf{P^T} && (9)\\
&= (\mathbf{PP^T})\mathbf{D}(\mathbf{PP^T}) && (10)\\
&= (\mathbf{PP^{-1}})\mathbf{D}(\mathbf{PP^{-1}}) = \mathbb{I}\mathbf{D}\mathbb{I} && (11)\\
\mathbf{C_Y} &= \mathbf{D} && (12)
\end{aligned}$$

donde puede apreciarse que la elección de $\mathbf{P}$ diagonaliza a $\mathbf{C_Y}$. Dentro de $\mathbf{P}$ y $\mathbf{C_Y}$ se encuentran los resultados de PCA:

- Los componentes principales son los eigenvectores de $\mathbf{C_X} = \frac{1}{n}\mathbf{XX}^T$.
- El i-ésimo valor diagonal de $\mathbf{C_Y}$ es la varianza de $\mathbf{X}$ a lo largo de $\mathbf{p_i}$

Los pasos para encontrar la solución a PCA para un conjunto de datos son mostrados en el Algoritmo 2.

3.1. Obtención y sustracción de la media de los datos

Se debe obtener la media y sustraerla a los datos por cada dimensión. Esta diferencia es utilizada para calcular la matriz de covarianza.

$$\mu = \sum_{i=1}^{n} x_i - \overline{x}$$

[1] Una matriz A es ortogonal si $AA^T = \mathbb{I}$

Algoritmo 2 Algoritmo para PCA a través de eigenvectores

1: Obtener y sustraer la media de los datos.
2: Calcular la matriz de covarianza Σ .
3: Calcular los eigenvectores para Σ .
4: Selección de componentes.
5: Formación del nuevo conjunto de datos.

3.2. Cálculo de la matriz de covarianza

Como ha sido descrito, la matriz de covarianza permite medir la redundancia y varianza de los datos. Es calculada a través de:

$$\Sigma = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)}$$

3.3. Cálculo de eigenvectores

El autovector, vector propio o eigenvector $\mathbf{v}$ de una matriz $\mathbf{M}$ de $(n \times n)$ es una matriz de $(n \times 1)$ tal que al multiplicarse por un elemento v obtiene un múltiplo de sí mismo. Esto se expresa como $A\mathbf{v} = \lambda v$, donde λ es un valor escalar real que recibe el nombre de autovalor, valor propio o eigenvalor.

A menudo, una transformación queda completamente determinada por sus vectores propios y valores propios. Entonces, su importancia radica en que pueden explicar la transformación que PCA realiza para encontrar la nueva base de los datos.

La mayoría de las herramientas computacionales para análisis de datos cuentan con alguna implementación para obtener eigenvalores y eigenvectores.

3.4. Selección de componentes

Los eigenvectores con los valores más altos son los componentes principales del conjunto de datos. Por lo tanto, los eigenvectores habrán de ordenarse de forma descendente (considerando el eigenvalor asociado a cada uno de ellos) para obtener los componentes por orden de importancia.

Existen distintos mecanismos para decidir la cantidad de eigenvectores a seleccionar, algunos indican seleccionar aquellos que describan a los datos al menos en un porcentaje determinado, otros cuando la separación entre dos eigenvectores consecutivos es grande.

A partir de los m eigenvectores seleccionados se forma la matriz de transformación $\mathbf{P}$:

$$P = \{\text{eig}_1, \text{eig}_2, \ldots, \text{eig}_m\}$$

3.5. Formación del nuevo conjunto de datos

Una vez seleccionados los eigenvectores a considerar, se debe crear el nuevo conjunto de datos. Esto es realizable a través de:

$$\text{Datos finales} = \mathbf{P}^T\mathbf{X}$$

4. Conclusiones

Se ha presentado una introducción a la reducción de dimensionalidad y una clasificación de las técnicas relacionadas. Asimismo, se ha presentado una descripción del fundamento matemático y comportamiento del PCA como herramienta no supervisada para la generación de las características más representativas de un conjunto de datos.