Generalized mean for robust principal component analysis

Erick S. Alvarez Valencia

7 de junio de 2018

Autores del paper

Jiyong Oh, Nojun Kwak

Contenido

Algoritmo clásico: PCA

Motivación

Se tiene un conjunto de m datos $X = (X_1, X_2, ..., X_m)$ los cuales viven en R^n y con matriz de covarianza Σ.

Suponiendo media cero para dicho conjunto de datos, se busca una matriz $W \in R^{n \times m}$ con $m < n$ tal que al realizar la proyección $W^T X$ se preserve la máxima cantidad de variabilidad en la misma.

Algoritmo clásico: PCA

Sea $Y = (Y_1, Y_2, ..., Y_m)$ donde $Y_i = W^T X_i$ las proyecciones del conjunto de datos con respecto a la matriz W. El problema de optimización se define como:

$$\begin{aligned} &\arg\max_{W} \ tr(W^T \Sigma W) \\ &s.a. \ W^T W = I \end{aligned} \tag{1}$$

Donde $\Sigma = \frac{1}{m} \sum_{i=1}^{m} X_i X_i^T$. Lo anterior es equivalente a

$$J_{L_2}(W) = \frac{1}{m} \sum_{i=1}^{m} ||X_i - WW^T X_i||_2^2 \tag{2}$$

Algoritmo clásico: PCA

Se puede mostrar que la solución del problema de optimización anterior es igual a

$$\Sigma v = \lambda v \tag{3}$$

Donde v es el eigenvector asociado al eigenvalor más grande de la matriz de covarianza. ¡Los eigenvalores explican la variabilidad!

Algoritmo clásico: PCA

Aunque PCA es muy simple y un buen método de reducción de dimensión, cuenta con fallos y uno de ellos es el hecho de que es sensible ante datos atípicos. El problema recae al hecho de que estos datos generan una gran variabilidad en el conjunto de datosy PCA puede dar soluciones no deseadas, y esto ocurre porque la función de optimización está formulada en base el error cuadrático medio y este a su vez usa la norma L_2.

Media generalizada

La media generalizada se define como:

Definición GM

Sea $p \neq 0$ y sea $A = (a_1, a_2, ..., a_m)^T$ un conjunto de datos positivos.

$$M_p\{a_1, a_2, ..., a_m\} = (\frac{1}{m}\sum_{i=1}^{m} a_i^p)^{1/p} \tag{4}$$

Casos especiales:

1. Si $p = 1$ se tiene la media aritmética.
2. Si $p \to 0$ se tiene la media geométrica.
3. Si $p = -1$ se tiene la media armónica.

Media generalizada

Se puede demostrar que:

$$\sum_{i=1}^{m} a_i^p = b_1 a_1 + b_2 a_2 + ... + b_m a_m \tag{5}$$

Donde $b_i = a_i^{p-1}$, $i = 1, 2, ..., m$.

Media muestral usando GM

PCA asume que los datos tienen media cero. La media muestral se puede formar como un problema de mínimos cuadrados.

$$m_s = \arg\min_{m} \frac{1}{M} \sum_{i=1}^{M} ||x_i - m||_2^2 \tag{6}$$

Como se usa la norma L_2 la media muestral es sensible ante datos atípicos.

Media muestral usando GM

Propuesta: Usar GM en la función de costo anterior para obtener una versión más robusta.

$$m_g = \arg\min_m \left(\frac{1}{M}\sum_{i=1}^{M}(||x_i - m||_2^2)^p\right)^{1/p},\ p > 0 \tag{7}$$

Como $p > 0$ lo anterior es igual a

$$m_g = \arg\min_m \sum_{i=1}^{M}(||x_i - m||_2^2)^p \tag{8}$$

Media muestral usando GM

Podemos utilizar la descomposición de GM como combinación lineal

$$\sum_{i=1}^{M}(||x_i - m||_2^2)^p \approx \sum_{i=1}^{M}\alpha_i^t||x_i - m||_2^2 \tag{9}$$

Donde $\alpha_i^t = (||x_i - m^t||_2^2)^{p-1}$. Para la ecuación anterior la aproximación se convertiría en igualdad cuando $m = m^t$. El siguiente paso es encontrar m^{t+1} tal que minimice α^t y en esta caso se puede derivar e igualar a cero, y con un poco de álgebra se obtiene:

$$m^{t+1} = \frac{1}{\sum_{j=1}^{M}\alpha_j^t}\sum_{i=1}^{M}\alpha_i^t x_i \tag{10}$$

Con esto se propone un método de optimización basado en el algoritmo EM.

Media muestral usando GM

Algorithm 1 Media generalizada

1: **procedure** GENERALIZEDMEAN(X, p)
2: t $\leftarrow$ 0.
3: $m^t \leftarrow m_s$.
4: **while** No convergencia **do**
5: **Aproximación** : Calcular $\alpha_1^t, \alpha_2^t, ..., \alpha_m^t$ usando (9).
6: **Minimización** : Usando las alfas calculadas en el paso anterior, calcular m^{t+1} con (10).
7: t $\leftarrow$ t + 1.
8: **end while**
9: **return** $m_g = m^t$.
10: **end procedure**

PCA robusto usando GM

Retomando PCA, tenemos que el error de reconstrucción generado por las proyecciones se calcula como

$$e(W) = \hat{X}^T \hat{X} - \hat{X}^T W W^T \hat{X} \tag{11}$$

Donde $\hat{X} = X - m$, es decir, los datos contienen media cero. Podemos aplicar en concepto de media generalida a la ecuación anterior para definir el nuevo problema de optimización, obteniendo:

$$W_g = \underset{W^T W = 1}{\arg\min} \left(\frac{1}{M} \sum_{i=1}^{M} [e_i(W)]^p\right)^{1/p}, \ p > 0 \tag{12}$$

Como $p > 0$ lo anterior es equivalente al siguiente problema

$$W_g = \underset{W^T W = 1}{\arg\min} \sum_{i=1}^{M} [e_i(W)]^p, \tag{13}$$

PCA robusto usando GM

Aplicando la descomposición de la media generalizada mostrada anteriormente tenemos

$$\sum_{i=1}^{M} e_i(W)^p \approx \sum_{i=1}^{M} \beta_i^t e_i(W)^p \tag{14}$$

Donde $\beta_i^t = [e_i(W^t)]^{p-1}$. Esta aproximación se vuelve exacta cuando $W = W^t$. Ahora para calcular W^{t+1} podemos usar la otra versión de la función objetivo de PCA

$$\begin{aligned} W^{t+1} = \arg\max_{W} \; & tr(W^T \Sigma_\beta^t W) \\ & s.a. \; W^T W = I \end{aligned} \tag{15}$$

Donde $\Sigma_\beta^t = \sum_{i=1}^{m} \beta_i^t \hat{x}_i \hat{x}_i{}^T$.

PCA robusto usando GM

Algorithm 2 PCA robusto con media generalizada

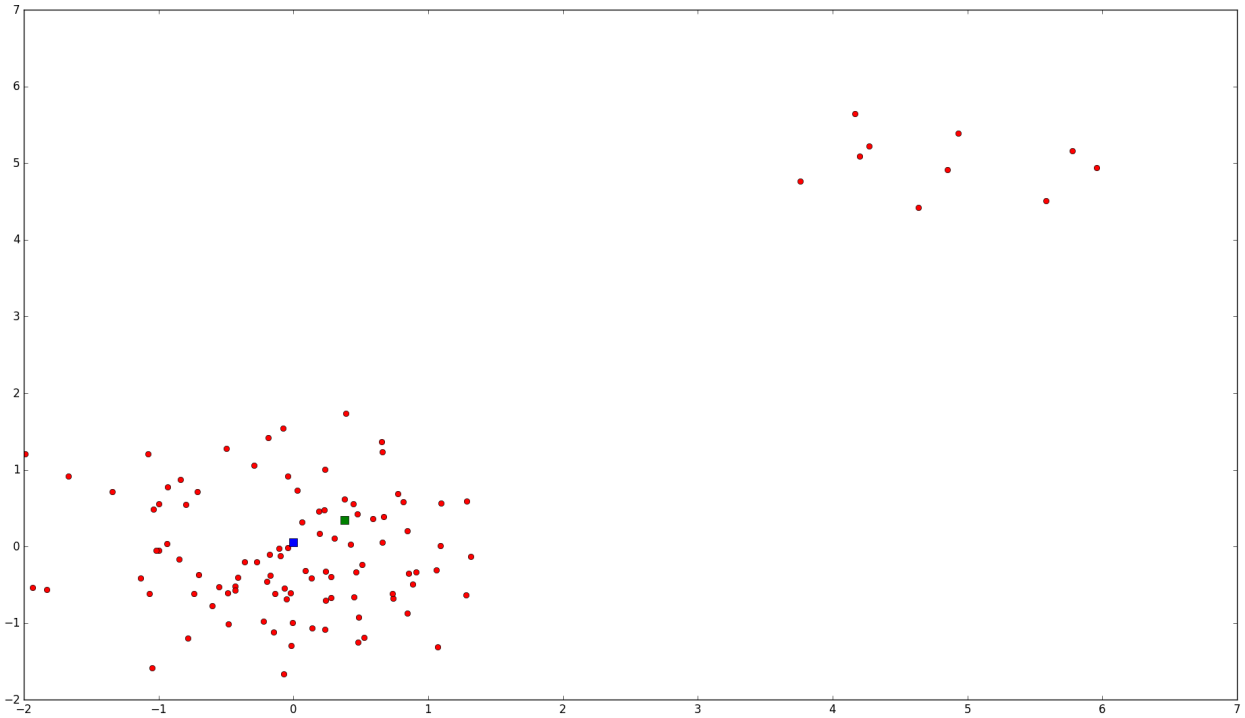
procedure RobustPCA(X, m_g, m, p)
    t $\leftarrow$ 0.
    $\hat{X} \leftarrow X - m_g$.
    $W^t \leftarrow W_{PCA}$.
    while No convergencia **do**
        Aproximación : Fijando W^t, calcular $\beta_1^t, \beta_2^t, ..., \beta_m^t$ usando la ecuación (14).
        Optimización : Usando las betas calculadas en el paso anterior, calcular W^{t+1} resolviendo el problema de eigenvalores-eigenvectores descrito en (15).
        t $\leftarrow$ t + 1.
    end while
    return $W_g = W^t$.
end procedure

Experimento 1

Se generaron 100 datos de una bivariada Gaussiana con media $\mu_i = (0,0)^T$ y matriz de covarianza $\Sigma_i = diag(0{,}5, 0{,}5)$ y se añeadieron 10 puntos atípicos de la misma distribución con media $\mu_o = (5,5)^T$ y matriz de covarianza $\Sigma_o = diag(0{,}3, 0{,}3)^T$. Posteriorme se obtuvo la media muestral y la media generalizada con $p = 0{,}2$.

Experimento 1

Resultados obtenidos

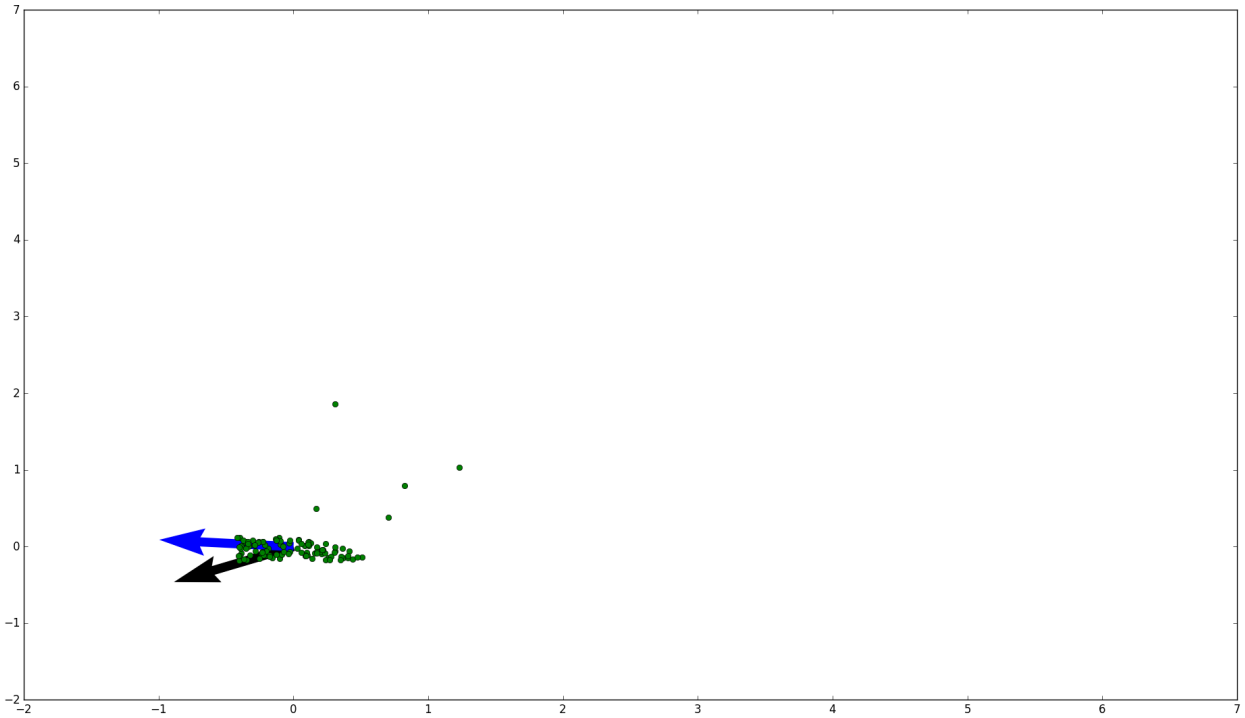


Experimento 2

Se generaron 100 datos $(x_i, y_i) \sim U(0,1)$ y se filtraron en una elipse centrada en el origen y con semiejes $(0{,}9^2, 0{,}1^2)$ y se añeadieron 5 puntos atípicos de la misma distribución con media $\mu_o = (1,1)^T$ y matriz de covarianza $\Sigma_o = diag(0{,}3, 0{,}3)^T$. Posteriorme se aplicó PCA y PCA robusto con $p = 0{,}3$.

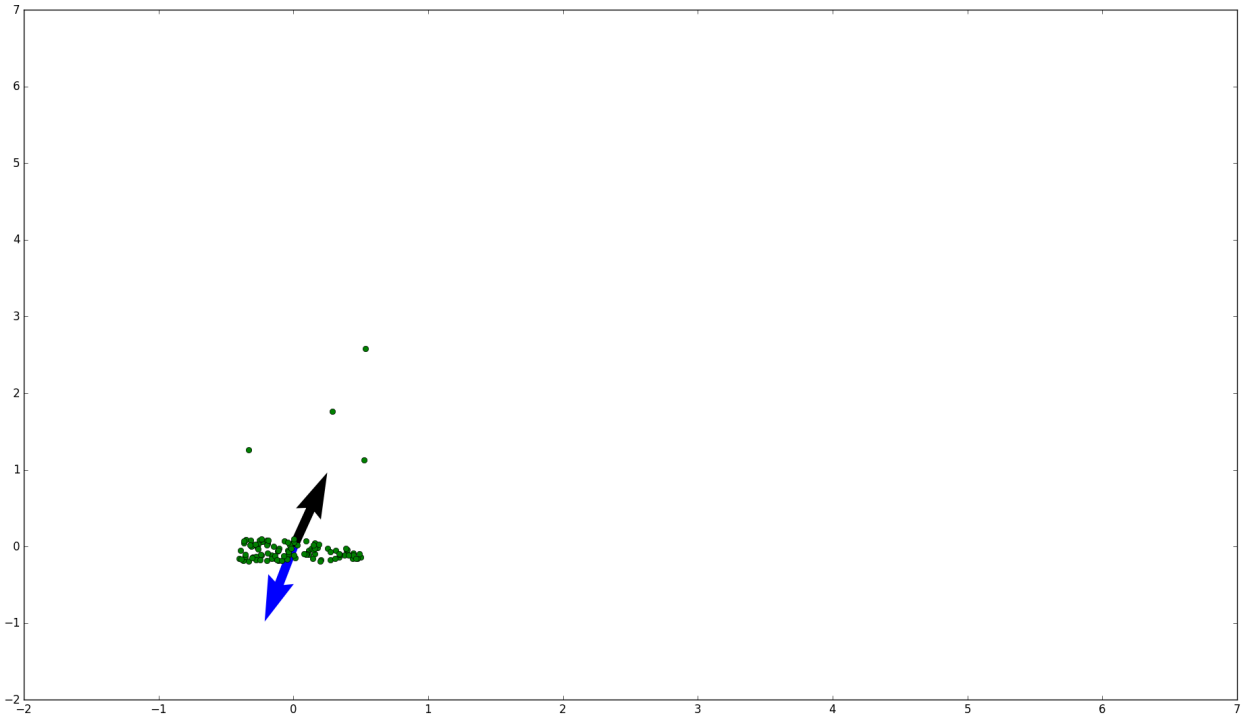
Experimento 2

Resultados obtenidos



Experimento 2

No todo es bonito : (

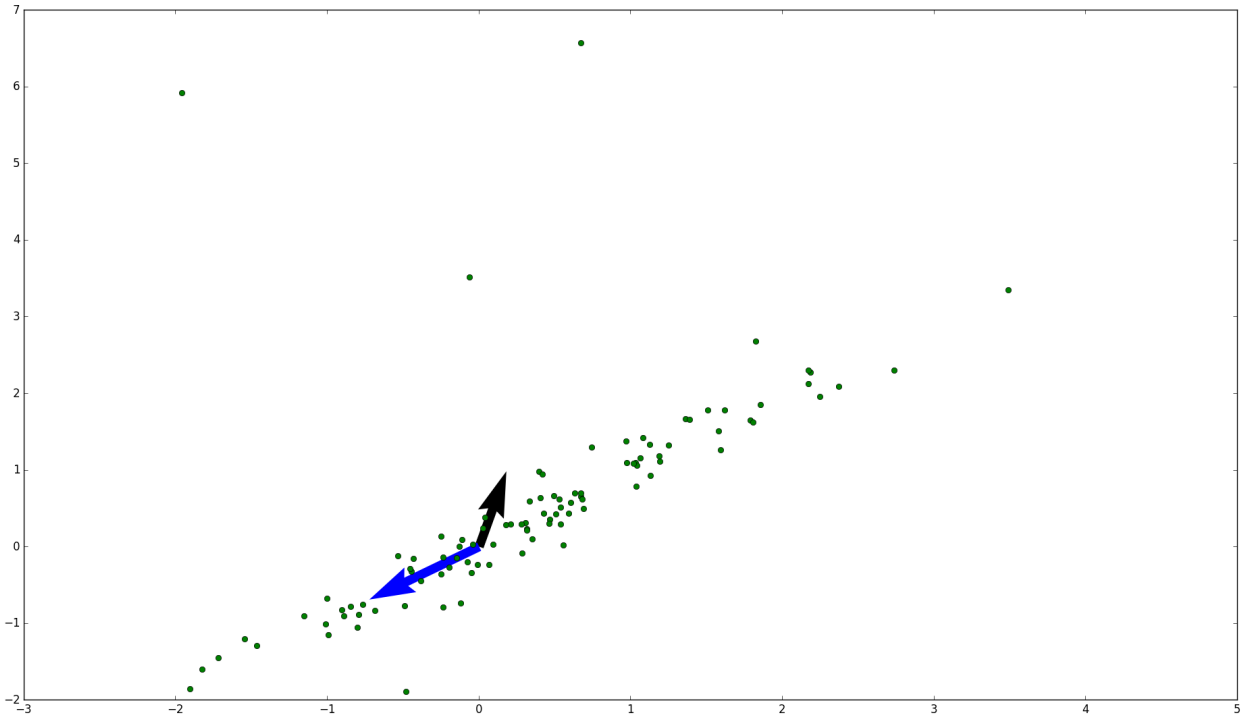


Experimento 3

Se generaron 100 datos usando una distribución univariada Gaussiana $x_i \sim N(0,1)$, $y_i = x_i + \epsilon_i$ con $\epsilon_i \sim N(0,0{,}5^2)$ para datos normales y $\epsilon_i \sim N(0,3^2)$ para datos atípicos. Posteriorme se aplicó PCA y PCA robusto con $p = 0{,}3$.

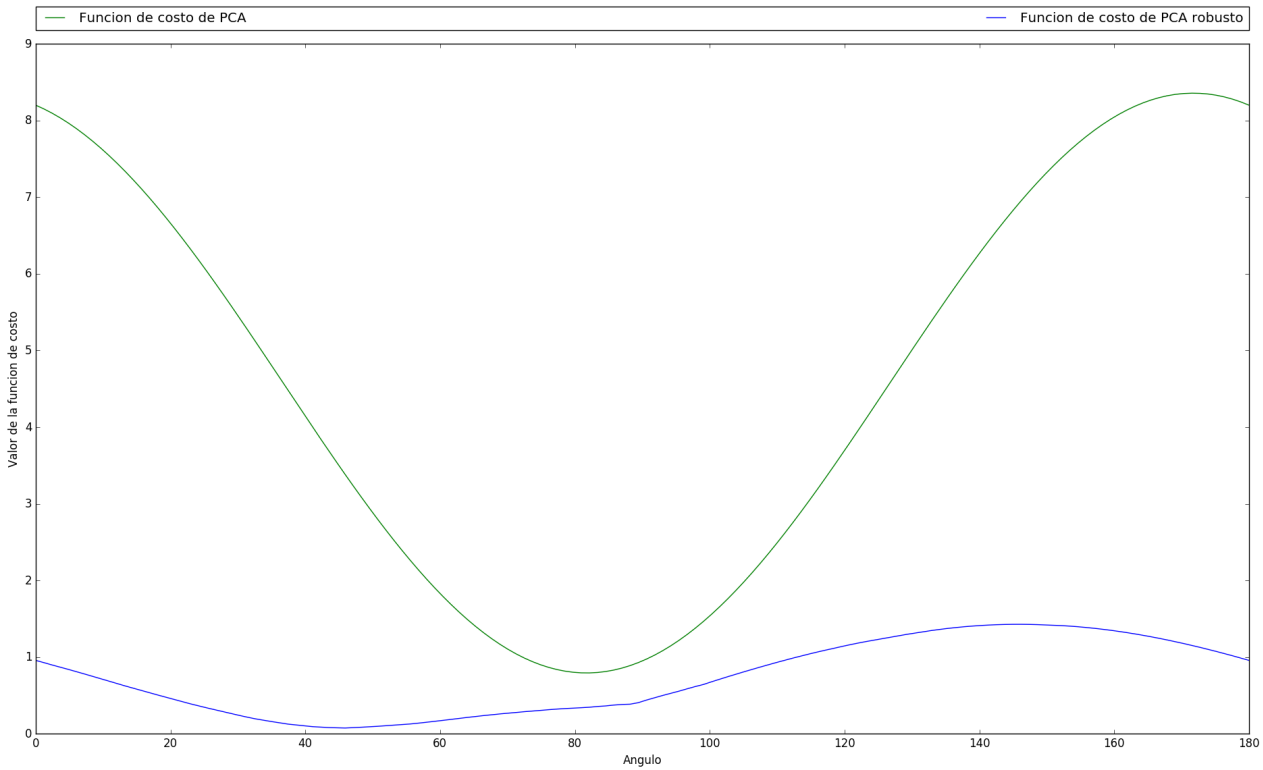
Experimento 3

Resultados obtenidos



Experimento 3

Gráfica que muestra la relación ángulo - función de costo para ambas versiones de PCA usando el conjunto de datos anterior.



Experimento 4

Se tomaron 300 muestras del dataset MNIST con los números 3, 8 y 9, luego 60 del mismo dataset con números diferentes para representar datos atípicos. Se normalizaron con norma L_1, se hizo reducción de dimension y se aplicó K-Means con 3 clusters.

Experimento 4

Cuadro: Resultados obtenidos del experimento anterior

m	PCA	PCA-GM
50	0.26	0.37
100	0.28	0.376
150	0.763	0.766
200	0.76	0.766
250	0.75	0.76
300	0.763	0.766

Conclusiones

1. Resultados favorables con respecto a PCA robusto.
2. Versión robusta en la mayoría de las ecuaciones superior a la versión clásica (al menos en los ejemplos vistos).
3. La convergencia del algoritmo no está asegurada aunque siempre se logró en los ejemplos probados.

Referencias

1. Jiyong Oh, Nojun Kwak. Generalized mean for robust principal component analysis, Pattern Recognition, Elsevier, 2016.