Generalized mean for robust principal component analysis

Erick Salvador Alvarez Valencia, *CIMAT A.C.*

***Index Terms*—Análisis de compontentes principales, datos atípicos, optimización.**

I. Introducción

En el presente reporte se describirá el análisis y la implementación del proyecto para la materia de optimización, el cual esta basado en un paper llamado *Generalized mean for robust principal component analysis* desarrollado por Jiyong Oh y Nojun Kwak ambos de la Universidad Nacional de Seoul. En el paper se propone una versión robusta ante datos atípicos del algoritmo de reducción de dimensión PCA, en donde hacen uso de la media generalizada o *power mean* para atacar el problema de los datos atípicos.
Se desglosará un poco la teoría sobre el método, de la misma forma se mostrará los resultados de la implementación del mismo y sobre todo lo que se obtuvo al reproducir algunos de los experimentos que se proponen en el paper.

mds
Junio 06, 2018

II. Análisis de componentes principales

La motivación de este proyecto fue la de trabajar con el clásico algoritmo de reducción de dimensión PCA o análisis de componentes principales pero en una versión robusta ante datos atípicos. Previo a explicar el método propuesto por el paper se asignará una sección para detallar un poco el método clasico el cual es el PCA.

PCA es probablemente la más antigua y más conocida de las técnicas de análisis multivariado, fue introducido por Pearson en 1901 y desarrollado de manera independiente por Hotelling alrededor de 1933. La idea central de PCA es reducir la dimensionalidad (a través de una proyección) de un conjunto de datos en el cual existen un gran número de variables interrelacionadas, manteniendo lo más que sea posible de la información presente en los datos. Por definición en PCA, se mide el grado de interés de una dirección a través de la variabilidad de los datos al ser proyectados sobre esta misma.

Dado un vector de datos $X = (X_1, X_2, ..., X_m)$ los cuales viven en $\mathbb{R}^n$ con matriz de covarianza Σ. Asumiendo sin pérdida de generalidad que los datos tienen media cero, estamos interesados en el problema de encontrar un subespacio $W \in \mathbb{R}^{nxm}$ con $m < n$ tal que al proyectar nuestro conjunto de datos en el mismo, se preserve la mayor cantidad de variabilidad de los mismos. Sea $Y = (y_1, y_2, ..., y_m)$ donde $y_i = W^T x_i$ proyecciones de los datos con respecto a la matriz W, el problema de optimización de PCA se define como:

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

Donde Σ es la matriz de covarianza de los datos y esta se puede calcular de manera muestral $\Sigma = \frac{1}{m}\sum_{i=1}^{m} X_i X_i^T$. La matriz de proyección W también se puede encontrar como la suma de los errores de proyección de los datos, y para ello se minimiza la siguiente función de costo:

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

Los dos problemas de optimización son equivalentes y se puede demostrar que la solución consiste en obtener los primeros eigenvectores asociados a los eigenvalores más grandes de la matriz de covarianza. Como dicha matriz es simétrica y semidefinida positiva, los eigenvalores resultantes de la misma serán reales y no negativos, por lo cual nos es posible ordenarlos de mayor a menor y encontrar los eigenvectores asociados a ellos.
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, dichos datos están alejados del conjunto original y es válido suponer que existen debido a una probabilidad muy baja en la distribución que genera a los datos. El problema recae al hecho de que estos datos generan una gran variabilidad en el conjunto y 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. En la siguiente sección se muestra la propuesta del paper para tratar de antemano este problema.

III. PCA robusto usando la media generalizada

Primeramente definimos la media generalizada de la siguiente manera: Sea $p \neq 0$ la media generalizada o $power\ mean$ M_p para un conjunto de datos positivos se calcula como

$$M_p\{a_1, a_2, ..., a_m\} = (\frac{1}{m}\sum_{i=1}^{m} a_i^p)^{1/p} \tag{3}$$

Se puede mostrar que la media aritmética, la media geométrica y la media harmónica son casos especiales de esta media cuando $p = 1$, $p \to 0$ y $p = -1$, respectivamente.

Además es posible mostrar que el mayor y el menor valor del conjunto se pueden obtener cuando $p \to \infty$ y $p \to -\infty$. Nota que cuando p decrece (incrementa), esta media es mayormente afectada por números pequeños (grandes), esto es el punto clave para tratar a los datos atípicos de un conjunto.
Se puede demostrar que la media generalizada de un conjunto de números positivos se puede expresar como una combinación lineal no negativa de los elementos del conjunto, tal como se muestra a continuación

$$\sum_{i=1}^{m} a_i^p = b_1 a_1 + b_2 a_2 + ... + b_m a_m \tag{4}$$

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

III-A. Media muestral generalizada

Si recordamos en la sección anterior PCA asume que los datos contienen media cero, es decir, están centrados por su media muestral. Ahora, el procedimiento de obtener la media muestral de un conjunto de datos puede verse como un problema de mínimos cuadrados.

$$m_s = \arg\min_m \frac{1}{M} \sum_{i=1}^{M} ||x_i - m||_2^2 \tag{5}$$

Y desde este punto datos atípicos influyen en el cálculo de la media y estos provocan que la misma se traslade más de lo que debería y por lo tanto, la media puede no quedar centrada en la nube de puntos de interés. Para esto se propone un algoritmo basado en la media generalizada para obtener una versión robusta de la media muestral, en dicho algoritmo la función de costo a obtimizar es la siguiente:

$$m_g = \arg\min_m \frac{1}{M} \sum_{i=1}^{M} (||x_i - m||_2^2)^p,\ p > 0 \tag{6}$$

Lo que se menciona en el paper en este punto es que se puede aplicar un algoritmo basado en gradiente para minimizar este problema, pero este enfoque resultaría lento para conjuntos de datos grandes y por lo tanto se propone un algoritmo basado en el algoritmo EM y en la descomposición de la media generalizada mostrada en la ecuación (4), la cual aplicándola a la función de optimización descrita en (6) quedaría de la siguiente manera:

$$\sum_{i=1}^{M} (||x_i - m||_2^2)^p \approx \sum_{i=1}^{M} \alpha_i^t ||x_i - m||_2^2 \tag{7}$$

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:

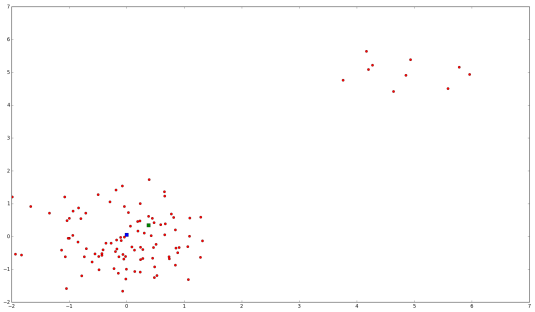
$$m^{t+1} = \frac{1}{\sum_{j=1}^{M} \alpha_j^t} \sum_{i=1}^{M} \alpha_i^t x_i \tag{8}$$

Este paso de actualización y el de calcular las alfas se hacen de manera separada en base a como trabaja el algoritmo EM y por lo tanto nos queda el siguiente algoritmo:

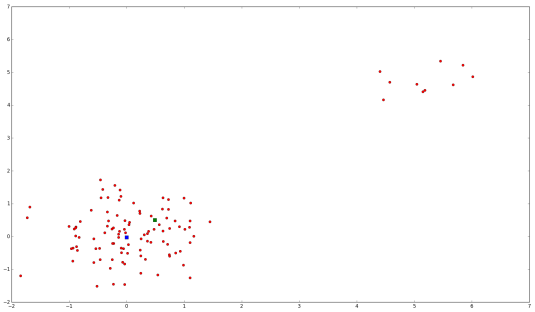
Algorithm 1 Media generalizada

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

Usando el algoritmo se reprodujo el primer experimento tratado en el paper, el cual consiste en generar un conjunto de 100 datos que viven en $\mathbb{R}^2$ usando una distribución bivariada gaussiana con media cero y matriz de covarianza $\Sigma_i = diag(0{,}5, 0{,}5)$, posteriormente a este conjunto se añaden una serie de 10 datos atípicos que de igual manera son generados con una distribución bivariada Gaussiana con media $m = (5, 5)^T$ y matriz de covarianza $\Sigma_o = diag(0{,}3, 0{,}3)$. El algoritmo de media generalizada fue aplicado a este conjunto de datos con $p = 0{,}1$ y $p = 0{,}2$, los resultados son mostrados a continuación.



(a) Figura 1. Resultados del aplicación del algoritmo de la media generalizada al conjunto de puntos descrito anteriormente y usando p = 0.1.



(b) Figura 2. Resultados del aplicación del algoritmo de la media generalizada al conjunto de puntos descrito anteriormente y usando p = 0.2.

En las gráficas anteriores podemos apreciar los resultados obtenidos al aplicar el algoritmo de la media generalizada al conjunto de datos explicando anteriormente, la Figura 1 muestra los resultados con p = 0.1 y la Figura 2 muestra los resultados con p = 0.2. En ambas imágenes se muestran los puntos calculados tanto con la media muestral (punto verde) como con la media generalizada (punto azul) y vemos que esta última no es afectada por los datos atípicos como pasa con la media muestral, por lo tanto vemos que este algoritmo funciona muy bien.

III-B. PCA robusto usando la media generalizada

Retomando el algoritmo de análisis de componentes principales, podemos ahora incluir la media generalizada en el problema mismo para tratar a los datos atípicos. Tenemos que el error de reconstrucción generado por las proyecciones se calcula como

$$e(W) = \hat{X}^T\hat{X} - \hat{X}^T W W^T \hat{X} \tag{9}$$

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:

$$W_g = \underset{W^TW=I}{\arg\min}\ (\frac{1}{m}\sum_{i=1}^{m}[e_i(W)]^p)^{1/p},\ p>0 \tag{10}$$

Lo cual es equivalente al siguiente problema

$$W_g = \underset{W^TW=I}{\arg\min}\ \sum_{i=1}^{m}[e_i(W)]^p,\ p>0 \tag{11}$$

Aplicando la descomposición de la media generalizada mostrada en (4) tenemos

$$\sum_{i=1}^{m} e_i(W)^p \approx \sum_{i=1}^{m}\beta_i^t e_i(W)^p \tag{12}$$

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

$$W^{t+1} = \underset{W}{\arg\max}\ tr(W^T\Sigma_\beta^t W) \quad s.a. W^TW = 1 \tag{13}$$

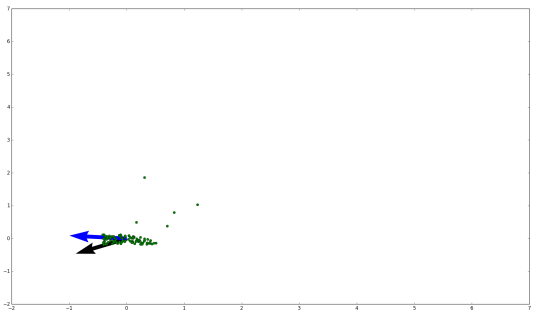
Donde $\Sigma_\beta^t = \sum_{i=1}^{m}\beta_i^t\hat{x}_i\hat{x}_i{}^T$.
De la misma forma el algoritmo de PCA robusto hace uso de la técnica empleada por el EM, de cálculo y actualización. Finalmente dicho algoritmo nos queda de la siguiente manera.

Algorithm 2 PCA robusto con media generalizada

procedure ROBUSTPCA(X, m_g, m, p)
  t ← 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 (12).
    Optimización : Usando las betas calculadas en el paso anterior, calcular W^{t+1} resolviendo el problema de eigenvalores-eigenvectores descrito en (13).
    t ← t + 1.
  return $W_g = W^t$.

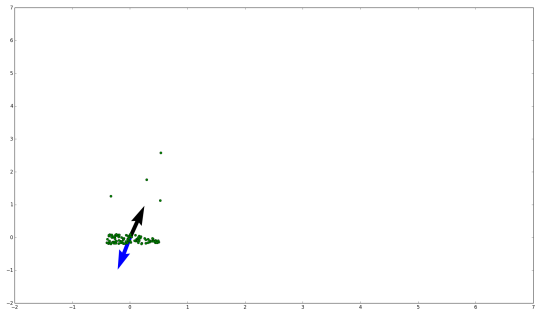
Podemos ver que en el algoritmo anterior Σ_β^t es la matriz de covarianzas pesada por los β_i^t, como dichos betas son positivos, la matriz sigue cumpliendo la propiedad de ser al menos semidefinida positiva. W^t es la matriz de proyecciones donde sus columnas son los eigenvectores que cumplen con ser las direcciones de máxima variabilidad, además de ser ortonormales. Y como nota importante, los datos inicialmente son centrados con la media generalizada implementada del algoritmo 1.

Una vez implementado el algoritmo anterior se procedió a realizar tres experimentos con el mismo para verificar si cumple con la propiedad de robustés ante los datos atípicos. El primer experimento no se encuentra en el paper, el cual consiste en generar un conjunto de puntos en el plano en forma de elipse y posteriormente añadir un pequeño conjunto de cinco datos atípicos alejados a cierta distancia, posteriormente a esto aplicar el PCA y el PCA robusto y comparar los resultados.
Los puntos no atípicos fueron generados con una distribución uniforme $U \sim (0,1)$ y posteriormente al ser generados se hizo un filtrado de los mismos para formar un conjunto en forma de elipse, dicha elipse esta centrada en $(0,0)^T$, su semieje de las abcisas era igual a $\sqrt{0{,}5}$, su semieje de las ordenadas era igual a $\sqrt{0{,}1}$. Los puntos atípicos fueron generados con una distribución bivariada Gaussiana con media $(1,1)^T$ y matriz de covarianza $\Sigma = diag(0{,}3, 0{,}3)$. Una vez generados los puntos se aplicaron ambos algoritmos en los mismos y se graficaron los resultados, los cuales se muestran a continuación.



(c) Figura 3. Resultados de PCA y PCA robusto al conjunto de puntos descrito para el experimento 2 usando p = 0.3.

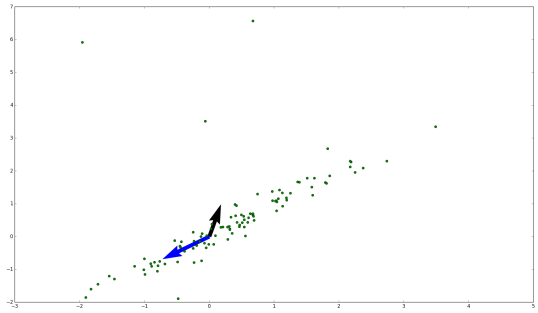
En la Figura anterior podemos ver el resultados del experimento 2, se pidió a los algoritmos que regresaran sólamente el primer vector propio el cual era el de máxima variabilidad, podemos ver que el vector resultante de PCA es el de color negro y el del PCA robusto es el de color azul. Se nota claramente que los datos atípicos no afectaron al resultado de PCA robusto mientras que el resultado de PCA si se dejó influenciar por los mismos ya que este trata de apuntar en dirección del elipse pero también trata de cubrir la variabilidad impuesta por los datos atípicos. Repitiendo este experimento muchas veces se llegó a casos donde el resultado de PCA robusto no fue bueno como se espera, a continuación se muestra la gráfica de ese caso.



(d) Figura 3. Resultados de PCA y PCA robusto al conjunto de puntos descrito para el experimento 2 usando p = 0.3.

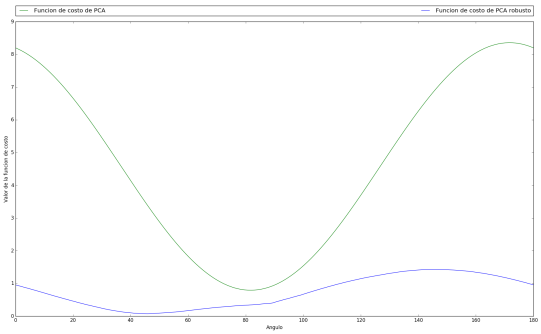
Se puede ver que ambas direcciones se vieron fuertemente influenciadas por los datos atípicos ya que apuntan hacia ellos y casi son paralelas. Pero cabe recalcar que se tuvo que repetir el experimento una gran cantidad de veces para llegar a este resultado.

El tercer experimento es propuesto por el paper y consiste en lo mismo que el anterior, se generó un conjunto de 110 puntos en un espacio de dimensión 2 (100 inliers y 10 outliers) donde $x_i \sim N(0,1)$ y $y_i = x_i + \epsilon_i$ donde $\epsilon_i \sim N(0,0{,}5^2)$ si era un punto normal y $\epsilon_i \sim N(0,3^2)$ si el punto era atípico. Posteriormente se aplicó tanto PCA como la versión robusta y se graficaron los resultados los cuales se muestra a continuación.



(e) Figura 4. Resultados de PCA y PCA robusto al conjunto de puntos descrito para el experimento 3 usando p = 0.3.

Se puede notar que al igual que el experimento anterior, la versión clásica de PCA se deja llevar por la variabilidad de los datos atípicos y por ende genera una dirección que apunta hacia ellos, mientras que en la versión robusta no pasa esto y la dirección generada es la que genera la variabilidad del conjunto normal y como si no existiaeran los puntos alejados. Si reproducimos este experimento con PCA clásico pero quitando los datos atípicos el primer eigenvector que retorna el método es aproximadamente $(cos(45°), sin(45°))^T$, los resultados de PCA robusto con datos atípicos para el experimento anterior fueron $(-0{,}724, -0{,}689)^T$ lo cual es aproximado a $-1 * (cos(45°), sin(45°))^T$, es decir, la solución anterior pero en una dirección paralela, mientras que PCA clásico ofreció una solución aproximada a $(cos(60°), sin(60°))^T$. Para terminar este experimento se generó una gráfica la cual en el eje X correspondía a un ángulo, en general $X \in [0, 180]$ y el eje Y correspondía al valor de la función de costo de PCA y de PCA robusto dándo como parámetro $W = (cos(X°), sin(X°))^T$ y como conjunto de puntos los del mismo ejercicio. A continuación se muestra la gráfica generada.



(f) Figura 5. Valor de las funciones de costo de PCA y PCA robusto con los datos del experimento 3.

En la Figura 5 se muestran las funciones de costo antes mencionadas, en general se puede ver que la gráfica verde corresponde a la función de costo de PCA y la azul corresponde a la de PCA robusto. Lo que se quiere mostrar aquí es que al menos para este conjunto de datos la solución de la función objetivo de PCA siempre acota superiormente a la de PCA robusto con respecto a cualquier ángulo de entrada.

Para el último experimento se trabajó con el conjunto de imágenes MNIST, la idea fue seleccionar un conjunto de números que fungieran como el conjunto normal y a esto añadirle un pequeño grupo de imágenes que tuvieran números distintos a los del conjunto original tal que fungieran como datos atípicos.
En general, se escogieron 300 imágenes que contuvieran los números 3, 8 y 9, y 60 adicionales con números diferentes a los anteriores, esto con el fin de ser la parte típica del problema. Posteriormente se aplicaron las dos versiones de PCA (clásica y robusta) con diferentes valores de n el cual era la dimensión del espacio proyectado y un valor de p el cual fue 0.4. Como las imágenes del conjunto MNIST son de dimensión 28x28 (784 si están vectorizadas), lo que se

quiere es ir trabajando con proyecciones muy pequeñas e irlas incrementando, por lo que la dimensión de las proyecciones trabajadas fue $n = \{50, 100, 150, 200, 250, 300\}$ con ambas versiones de PCA y hay que denotar que previo al proceso de proyectar con respecto a W se normalizaron los datos con recpecto a la norma L_1. El último paso de este experimento fue hacer clustering a los datos proyectados usando el método de K-Means con 3 clusters y para los clusters iniciales del método se eligieron las 3 observaciones más alejadas entre sí con respecto a su media L_2 y finalmente se midió la precisión de los grupos resultantes que hizo K-Means con los datos proyectados para PCA y PCA robusto. A continuación se muestran dichos resultados:

Cuadro I: Resultados obtenidos del experimento 4.

n	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

En la tabla anterior se muestra la precisión generada por K-Means utilizando diferentes valores de n y un valor de $p = 0{,}4$. Primeramente podemos notar que para una proyección en una dimensión de $n = 50$ se obtiene una predicción en K-Means muy baja tanto si se usa PCA como PCA robusto, aunque se puede destacar que la versión robusta supera por poco a la clásfica. En base a como va aumentando la dimensión de las proyecciones podemos ver que la precisión del clustering también va aumentando para ambas formas de proyección, y a excepción del caso $n = 100$ vemos que la precisión con respecto a ambos métodos de PCA es muy parecida pero siempre la versión robusta supera a la clásica. Para $n = 100$ si tenemos una clara mejora con respecto a la versión robusta y en general para pocas dimensiones es algo que esperaríamos en el sentido de clasificación que K-Means tenga una notable mejora en su precisión con PCA robusto, esto porque para este número de dimensiones, la variabilidad de la proyección es muy poca comparado a la del conjunto original y lo que estamos viendo es que la versión robusta pudo capturar mejor esta variabilidad que la versión clásica. Si vemos desde $n = 150$ los resultados se empiezan a parecer mucho y es porque pasa lo contrario, para este punto el número de componentes principales ya explican un alto porcentaje de la variabilidad (entre 87 % y 90 %) por lo que no debería haber una mejora muy significativa como lo es con dimensiones bajas.

IV. Conclusión

En el presente reporte se mostró la implementación de una propuesta de mejora hacia el método de análisis de componenetes principales basado en el uso de la media generalizada.

Se pudo mostrar los algoritmos derivados de la media generalizada y usando la estrategia que usa el algoritmo EM tanto para la media como para PCA, además de los resultados de la implementación de los mismos en varios experimentos que usaban datos sintéticos y uno que trabajaba con el conjunto de datos MNIST, y en todos las pruebas se encontraron buenos resultados y para todas se tuvo una mejora de la versión robusta de PCA con respecto a la clásica pero de igual manera este contiene algunas desventajas las cuales son:

1. No se cuenta con una solución cerrada para la resolución de este problema por lo tanto se requiere de un método iterativo.
2. No se tiene asegurada la convergencia del algoritmo propuesto.
3. Para ciertos conjuntos de datos, el método no necesariamente ofrecerá soluciones robustas, aunque en el peor de los casos este ofrecerá la solución de PCA clásico.

Finalmente, aunque el algoritmo cuenta con dichas desventajas, se puede decir que esta es una gran propuesta basada en una idea sencilla con buenosnos resultados.

Referencias

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