Análisis Sparse de Tensores Multidimensionales como alternativa al Análisis de Componentes Principales clásico: ventajas e inconvenientes

N. González[1], M. P. Galindo[2] y A. B. Nieto[2]

[1] *Departamento de Estadística, Universidad de Salamanca, Campus Miguel de Unamuno. 37007 Salamanca, nerea_gonzalez_garcia@hotmail.com*

[2] *Departamento de Estadística, Universidad de Salamanca*

Resumen

El Análisis de Componentes Principales (PCA) (Jolliffe, 2002) es la técnica más utilizada del análisis estadístico multivariante. Permite, a partir de un conjunto de *p* variables relacionadas, extraer *q* variables latentes (conocidas como componentes principales, PCs), con *q* mucho menor que *p*, y con ellas conocer el comportamiento de la muestra en un espacio de baja dimensión. Sin embargo, esta técnica tiene un principal inconveniente: cada PC es una combinación lineal de todas las variables originales y esto dificulta su interpretación. Se analiza aquí una alternativa: Análisis de Componentes Principales Sparse (SPCA) (Zou, Hastie, & Tibshirani, 2006), introducida en la literatura desde hace años, pero no utilizada en la práctica hasta el momento. Gracias a la reformulación del PCA como un problema de minimización del error se facilita la modificación del algoritmo mediante la adición de la penalización Elastic net (Zou & Hastie, 2005). Esto dará lugar a vectores de cargas *sparse* (nulas), convirtiéndose el SPCA en un método de selección automática de variables características, que no conlleva la subjetividad de los investigadores. Así, cada componente será una combinación sólo de las variables realmente relevantes.

Habitualmente, los estudios recogen la información en matrices de dos vías. Sin embargo, el avance en los mecanismos de recolección de datos y el interés por recoger la mayor información posible, ha provocado que las bases de datos disponibles sean grandes y complejas. Es por esto que la información se presenta en matrices multidimensionales (Cichocki, Zdunek, Phan, & Amari, 2009; Kroonenberg, 2008). A través de los métodos pertinentes para su estudio (Tucker (Tucker, 1966, 1972), PARAFAC (Harshman, 1970)), se obtendrían conclusiones más eficientes que si esas matrices se estudiasen por separado. Ahora bien, estos métodos basan su fundamento teórico en el PCA. Se propone una nueva forma de enfocar las técnicas de factorización/descomposición de matrices multidimensionales fundamentándose en el Sparse PCA. Esto favorecerá su aplicación en áreas como genética, cuyos resultados con las técnicas clásicas no son óptimos.

Objetivos

El **objetivo global** del proyecto radica en la fundamentación de un método de análisis multivía con base en el Sparse PCA.

Los **objetivos directrices** del proyecto son:

i) Evidenciar las dificultades existentes en el análisis del tipo de matrices de datos actuales con las técnicas clásicas.
ii) Revisión bibliográfica de las principales vertientes en que se presenta el Sparse PCA, así como sus ventajas e inconvenientes.
iii) Desarrollo de la metodología para la óptima selección de los parámetros de regularización.
iv) Desarrollo teórico del modelo de Sparse PCA como técnica penalizada con una óptima selección de los parámetros de regularización, así como su implementación en software libre.
v) Mostrar el uso útil del sparse PCA a nuevos campos; por ejemplo, en áreas del ámbito social.
vi) Estudio de los tensores, definición propiedades, y resultados principales.
vii) Revisión bibliográfica de los principales métodos multivariantes de análisis de matrices multivía (factorización y descomposición de tensores multivía).
viii) Extensión y desarrollo del modelo sparse propuesto para el análisis de matrices multivía.
ix) Comprobación de las propiedades óptimas de los modelos con su aplicación a datos reales.

Desarrollo de la investigación

El PCA (Jolliffe, 2002) se ha convertido en la técnica más utilizada para conocer el comportamiento de una muestra. Es una técnica de extracción de características y de reducción de la dimensión de matrices de datos que permite la reproducción de los datos en un espacio de baja dimensión. Extrae q variables latentes incorrelacionadas (componentes principales, PCs) a partir de p variables originales correlacionadas, con $q<<p$. Ahora bien, esta técnica presenta dos deficiencias principales: su interpretación, debido a que cada PC es una combinación lineal de todas las variables originales, y su inconsistencia para datos de altas dimensiones.

Con la finalidad de solventar estas deficiencias se desarrolla el Sparse PCA, método para producir PCs modificadas con algunos coeficientes de la combinación lineal (cargas) nulos. El Sparse PCA es una técnica de selección de variables que ha recibido más atención en la última década y se han desarrollado varios algoritmos para la obtención de cargas *sparse* en base a diversos criterios (Ning-min & Jing, 2015).

Al igual que su antecesor PCA, el Sparse PCA cuenta con distintas formulaciones del problema. Principalmente, existen dos formas de lograr la nulidad en las cargas: mediante la restricción de las cargas o a través de su contracción. Dentro de los métodos que tratan de contraer los vectores de cargas, el principal es el SPCA presentado por Zou et al.

(2006). Formulan el problema del SPCA como un problema de optimización del tipo regresión penalizada con Elastic net (Zou & Hastie, 2005), sin embargo cuenta con una principal desventaja: la selección de los parámetros de regularización que restringen los vectores de cargas. En este trabajo se pretende aportar una buena base del método para el análisis de matrices bidimensionales con una óptima selección de los parámetros de regularización, pues una vez cimentado el método en dos vías estaremos en condiciones de presentar nuestra contribución principal.

Los métodos clásicos del análisis multivariante trabajan con matrices de dos vías; sin embargo, en diversas disciplinas, ha surgido la necesidad de manipular datos descritos en múltiples dimensiones denominados tensores (Kroonenberg, 2008). Al igual que en el caso bidimensional, el reemplazo de estos tensores por una aproximación de menor dimensión permite observar estructuras que a priori no podrían ser observadas. Actualmente, los trabajos de integración de matrices quedan recogidos en dos vertientes fundamentales aunque nosotros consideraremos la segunda: los métodos franceses y los anglosajones. De entre los métodos anglosajones, caracterizados por ajustar modelos que reproduzcan lo mejor posible los datos originales, destacan los modelos de Tucker (Tucker, 1966, 1972), el método PARAFAC (Harshman, 1970) y los métodos Tuckals (Kroonenberg & Leeuw, 1980). Estos métodos tienen su cimiento teórico en el PCA. Por ello, a través de la teoría algebraica de tensores, proponemos el desarrollo de un modelo de análisis de tensores basado en el Sparse PCA.

La metodología a utilizar será la siguiente:

a. Manifestación de la necesidad de desarrollo de nuevas técnicas para el análisis de bases de datos actuales.
b. Revisión bibliográfica de los métodos de Análisis de Componentes Principales Sparse en el caso bidimensional y multidimensional.
c. Desarrollo teórico del modelo de factorización/descomposición SPCA con para el análisis de matrices multidimensionales.
d. Implementación de software para su uso.

La presentación de esta unificación se llevará a cabo según el siguiente esquema:

Tarea 1. Presentación de resultados no óptimos de las técnicas clásicas.

Tarea 2. Revisión bibliográfica de los principales métodos multivariantes aplicados para el estudio de matrices de dos vías que mejoren la interpretación de los resultados obtenidos.

Tarea 3. Formulación y comparación de los diferentes enfoques de formulación del SPCA, sus ventajas e inconvenientes.

Tarea 4. Desarrollo del modelo teórico de Sparse PCA adecuado para su extensión al caso multidimensional y la correcta selección de los parámetros a estimar.

Tarea 5. Revisión bibliográfica de los principales métodos multivariantes aplicados para el estudio de matrices multivía.

Tarea 6. Desarrollo del modelo teórico de Sparse PCA para tensores multidimensionales.

Tarea 7. Implementación de interfaces gráficas para la utilización de las versiones sparse de los métodos desarrollados anteriormente.

Tarea 8. Aplicación a datos reales y simulados.

Resultados

Los resultados específicos de este estudio son:

- Presentar una solución óptima del problema habitual de componentes principales sparse que presentan la mayor parte de los distintos modelos propuestos para el Sparse PCA en dos vías: selección de parámetros de regularización.
- Proporcionar un modelo eficiente que proporcione el análisis de grandes matrices de datos en un tiempo óptimo.
- Extender este modelo al análisis de matrices en el que se consideran más de dos condiciones (matrices o tensores multivía).
- Implementación del software para su uso para matrices de dos vías y multivía.
- Demostrar las ventajas en la aplicación de datos de esta técnica frente a las opciones de las técnicas clásicas, mostrando la aplicación del Sparse PCA.

Conclusiones

El interés científico es claro ya que con el presente proyecto se desarrolla una nueva metodología que permitirá analizar matrices de datos complejas, cada vez más frecuentes y difícilmente analizables. La técnica del PCA es la más recurrida para el análisis de matrices, pero aquí aplicada presenta algunas deficiencias: puede ser que no proporcione interpretaciones factibles o bien que directamente no sea aplicable.

El Sparse PCA encaja a la perfección en una de las necesidades del momento: el análisis de grandes matrices de datos, en el que el principal objetivo no es seleccionar la mayor parte de información, sino seleccionar la información característica. Favorecemos así el estudio de conjuntos de datos que con técnicas clásicas no son óptimos. El desarrollo de una técnica como el Sparse PCA en el análisis de matrices multidimensionales permitiría solventar esta problemática, reduciendo además el tiempo requerido de estudio. La tecnología de recolección de datos avanza y, por ello, las técnicas tienen que avanzar con ellas.

El interés social de este proyecto radica en el ámbito en que se aplique. Permitiría que estudios biológicos, clínicos, informáticos,… de gran importancia obtuvieran resultados eficientes de manera rápida.

Referencias

Cichocki, A., Zdunek, R., Phan, A. H., & Amari, S. I. (2009). *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. Wiley. Retrieved from http://doi.org/10.1002/9780470747278

Harshman, R. a. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, *16*(10), 1– 84.

Jolliffe, I. (2002). Principal component analysis. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat06472/full

Kroonenberg, P. M. (2008). *Applied Multiway Data Analysis*. Wiley. Retrieved from http://doi.org/10.1111/j.1751-5823.2008.00062_6.x

Kroonenberg, P. M., & Leeuw, D. J. (1980). Principal Component Analysis of 3-Mode Data by Means of Alternating Least-Squares Algorithms. *Psychometrika*, *45*(1), 69–97.

Ning-min, S., & Jing, L. (2015). A Literature Survey on High-Dimensional Sparse Principal Component Analysis. Overview of PCA and Sparse PCA, *8*(6), 57–74.

Trendafilov, N. T. (2014). From simple structure to sparse components: A review. *Computational Statistics*, *29*(3-4), 431–454.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, *31*(3), 279–311. doi: 10.1007/BF02289464

Tucker, L. (1972). Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika, 37(1),* 3-27. doi: 10.1007/BF02291410

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*. doi: 10.1198/106186006X113430