APRENDIZAJE AUTOMATICO 1

Articulo científico:

Article (1965) _Principal Components Regression in Exploratory Statistical Research

Integrantes:

- Andres Felipe Borrero
- Carlos Alberto Martinez
- Nicolas Colmenares
- Yesid Castelblanco

Profesores:

- Santiago Ortiz
- Henry Velasco

Introducción

La regresión por componentes principales (PCR) es una herramienta fundamental en la investigación estadística exploratoria, especialmente en contextos donde la multicolinealidad y la alta dimensionalidad de los datos presentan desafíos significativos. Este ensayo se centrará en el artículo "Principal Components Regression in Exploratory Statistical Research", el cual destaca las innovaciones y ventajas de la PCR frente a los métodos tradicionales de regresión.

El artículo subraya cómo la PCR permite manejar datasets con alta multicolinealidad, facilitando la exploración de relaciones complejas entre variables en un espacio de datos simplificado. A diferencia de la regresión tradicional, que enfrenta dificultades al aislar el impacto individual de cada variable independiente, la PCR utiliza componentes principales ortogonales para resumir la información y reducir la dimensionalidad, mejorando así la estabilidad y la interpretabilidad del modelo.

Se destacan las ventajas computacionales de la PCR, como la reducción del esfuerzo computacional y la capacidad de procesar datos en paralelo, lo que la hace ideal para sistemas de alto rendimiento. Esto resalta la importancia de seleccionar cuidadosamente

los componentes principales, asegurándose de incluir solo aquellos que estén fuertemente relacionados con las variables dependientes para evitar una complejidad innecesaria en el modelo.

En conclusión, la PCR no solo ofrece una solución eficaz para problemas de multicolinealidad y alta dimensionalidad a través de la transformación de las variables originales en un nuevo conjunto de variables no correlacionadas (componentes principales) y utilizando estos componentes en el modelo de regresión, sino que también proporciona un marco flexible y robusto para la investigación estadística exploratoria.

Este ensayo explorará algunos detalles de los beneficios del PCR, limite en el número de componentes y para la inferencia, selección del mejor modelo PCR además de la selección de espacios según conocimiento de negocio vs PCA.

1. Beneficios de PCR

El análisis de componentes principales (PCA) es un método conocido desde hace muchos años y ha sido discutido por diversos autores como Anderson y Kendall. Es especialmente útil en situaciones donde no se tienen modelos claros o hipótesis definidas sobre las relaciones entre variables, siendo de gran valor en la investigación exploratoria. Utilizando técnicas como el PCA, los investigadores pueden identificar patrones en los datos que podrían ser cruciales para estudios en áreas donde los datos son tan complejos que los métodos estadísticos tradicionales no funcionan bien.

El PCA es especialmente útil cuando las variables independientes originales están muy correlacionadas (multicolinealidad) o cuando hay muchas variables explicativas. Al transformar un conjunto de variables en componentes principales, se facilita la exploración de sus relaciones con una variable dependiente.

1.1. Ventajas de la PCR

- **Manejo de Multicolinealidad:** Al convertir las variables originales en componentes principales que no están correlacionados entre sí, el PCR permite hacer análisis de regresión sin los problemas que causa la multicolinealidad. Esto resulta en estimaciones de los coeficientes que son más estables y confiables.
- **Reducción de Dimensionalidad**: Esto simplifica el modelo y facilita la interpretación, al tiempo que se conserva la mayor parte de la variabilidad de los datos originales. Esto es especialmente valioso en conjuntos de datos con un gran número de variables.
- **Máxima Información en Componentes:** Cada componente principal se crea para capturar la mayor cantidad de información posible de las variables originales y, al mismo tiempo, no estar correlacionado con los componentes anteriores. Esto

permite una representación más detallada y completa de la estructura de los datos, en comparación con otros métodos que no usan esta técnica.

1.2. Ventajas Computacionales

- **Reducción de la Carga Computacional:** Al transformar un conjunto de variables originales en un número reducido de componentes principales, el PCR disminuye la carga computacional necesaria para realizar análisis de regresión, generando así regresiones con menor complejidad computacional al trabajar con un set de variables de menor tamaño al dataset original. Esta reducción de la dimensionalidad reduce la carga computacional.
- **Exploración simplificada de modelos alternativos**: PCR permite la exploración de modelos de regresión a través de simplemente seleccionar subsets de los componentes principales sin necesidad de ejecutar el Análisis de Componentes en cada exploración. En métodos tradicionales puede ser necesario ejecutar la regresión para cada variación del modelo.
- **Manejo Eficiente de Grandes Conjuntos de Datos**: Al reducir la cantidad de variables y trabajar con menos componentes, se simplifica el proceso de cálculo. Esto permite hacer análisis más rápidos y con menos errores, en comparación con métodos que necesitan usar todas las variables originales.

1.3. Innovaciones del estudio

La principal innovación de las PCR está en la habilidad de tratar con datasets de alta multicolinealidad.

La habilidad previamente mencionada se da por la forma de resumir datos y reducir la dimensionalidad. Esto permite a los investigadores explorar relaciones complejas entre variables dentro de un espacio de datos simplificado.

Las comparaciones entre los resultados obtenidos mediante el método de regresión de componentes principales y los de la regresión clásica, mostrando la efectividad del enfoque propuesto en el análisis de datos complejos.

2. Limite en el número de componentes calculados

2.1. Entre más aumenta el número de PC, aumenta la varianza (riesgo de overfitting?)

En un PCA, un aumento en el número de componentes principales puede ir relacionado con un aumento en la varianza, sin embargo, esto puede llevar a un modelo que se ajusta

demasiado a los datos de entrenamiento, capturando ruido en lugar de patrones significativos, lo que incrementa el riesgo de overfitting. Si utilizamos demasiados componentes, la validación cruzada puede mostrar un buen rendimiento en los datos de entrenamiento, pero un rendimiento malo en el conjunto de datos de prueba, esta es una señal muy clara de overfitting. Es importante validar el modelo con un conjunto de datos independiente o mediante técnicas de validación cruzada. Esto ayuda a determinar si el modelo realmente se beneficia de la inclusión de más componentes o si está sufriendo de overfitting.

2.2. Entre más aumenta el número de PC, el R2 llega a un límite

En el contexto de PCR, el R2 tiende a alcanzar un límite a medida que se aumenta el número de componentes principales. Con cada componente principal que se agrega, el R2 normalmente aumenta y esto debido a que cada PC adicional puede generar más varianza, sin embargo, ese aumento se vuelve marginal después de cierto número de componentes ya que los primeros componentes principales capturan la mayoría de varianza significativa.

3. Selección del mejor modelo PCR

Es importante tener en cuenta el objetivo principal del modelo. Si el interés radica en mejorar la predicción y se cuenta con un gran número de variables, o si se ha detectado multicolinealidad durante el análisis exploratorio, el uso de componentes principales en la regresión puede ser una solución adecuada. Esto permite reducir tanto la cantidad de variables como la multicolinealidad, ya que los componentes principales son combinaciones lineales de las variables originales. Sin embargo, cuando se transforman los datos, no es posible revertir estos componentes a las variables originales, especialmente cuando se ha reducido la dimensionalidad. Además, a medida que se aumenta el número de componentes principales, se utilizan más combinaciones de las variables originales, lo que reduce la capacidad de interpretación del modelo. Por lo tanto, aunque el análisis de componentes principales es útil para abordar problemas como la multicolinealidad y la reducción de dimensiones, su uso debe ser considerado cuidadosamente. Si la interpretabilidad y la capacidad de inferencia del modelo son prioridades, se debe evaluar si la reducción dimensional a través de componentes principales es la mejor opción o si es preferible optar por otras técnicas que mantengan la relación directa con las variables originales.

4. Selección de espacios según conocimiento de negocio vs PCA

Es posible adaptar el concepto de componentes principales de acuerdo con el conocimiento del negocio. Por ejemplo, si se tiene un dataset cuyo objetivo es predecir el valor de una empresa y se dispone de variables como ingresos netos, intereses e impuestos, depreciación y amortización, no se podría graficar debido al número de variables. Sin embargo, se puede utilizar el análisis de componentes principales para reducir estas

variables a 1 o 2 dimensiones, facilitando su visualización. A pesar de esto, este enfoque puede dificultar la capacidad de inferencia. Afortunadamente, en el contexto del negocio existe una fórmula que combina estas variables: el EBITDA, que es la suma de todas ellas. Con lo anterior, es similar a usar un único componente principal, reduciendo de 5 a 1 variable explicativa, con el beneficio adicional de mantener la interpretabilidad del modelo.

Conclusiones

1. En casos donde se tengan muchas variables o se detecte multicolinealidad el PCA es muy útil para resolver ambos problemas. Sin embargo, se debe considerar que esto reduce la capacidad de inferencia del modelo, por lo que su uso busca mejorar en la predicción. En otras palabras, si te interesa que el valor predicho sea el más cercano a la realidad este método es el adecuado, pero si tu interés radica en identificar como influye cada variable predictora en el valor predicho, es buscar otras alternativas.

2. No existe una aproximación analítica para encontrar los mejores componentes principales, al igual que Elastic-Net, es necesario usar cross-validation para encontrar el mejor modelo.

3. Existen casos donde la combinación lineal de las variables existentes está dada por el negocio en la que se aplica, por ejemplo, una empresa con datos como sus ingresos netos, los intereses e impuestos pagados, depreciación y amortización, en lugar de usar cada valor por separado o tratar de hallar un PCA adecuado se puede usar el EBITDA que es la sumatoria de estos campos para pasar de 5 a 1 dimensión. Lo anterior, sigue siendo PCA con un solo componente principal, teniendo la ventaja que en este caso permite una inferencia mejor gracias a conocer la combinación de las variables.