INDEPENDENT COMPONENT ANALYSIS (ICA)

INTEGRANTES:

VIVIANA ANDREA GARCIA MONJE
VALERIA ROMERO PEREZ
YENIFER PATRICIA GUAJE NIÑO

INSTRUCTOR:

DAVID FRANCISCO BUSTOS USTA

SENA - BOOTCAMP DATA SCIENCE
SEPTIEMBRE 16 DE 2021

Análisis de Componentes Independientes (ICA)

Historia y Antecedentes

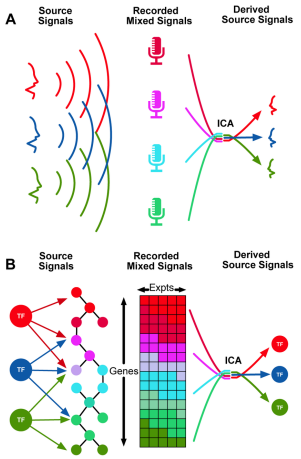
El primer marco general para el análisis de componentes independientes fue introducido por Jeanny Hérault y Bernard Ans en 1984,desarrollado por Christian Jutten en 1985 y 1986, y perfeccionado por Pierre Comon en 1991, y popularizado en su artículo de 1994. En 1995, Tony Bell y Terry Sejnowski introdujeron un algoritmo ICA rápido y eficiente basado en infomax , un principio introducido por Ralph Linsker en 1987.

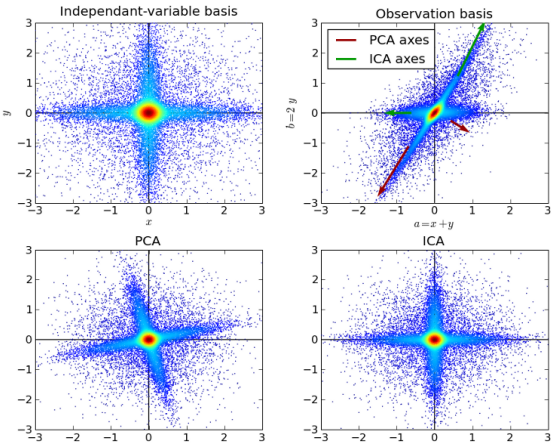
Hay muchos algoritmos disponibles en la literatura que hacen ICA. Uno de los más utilizados, incluso en aplicaciones industriales, es el algoritmo FastICA, desarrollado por Hyvärinen y Oja, que utiliza la curtosis como función de coste. Otros ejemplos están más bien relacionados con la separación ciega de fuentes donde se utiliza un enfoque más general. Por ejemplo, se puede descartar el supuesto de independencia y separar señales correlacionadas entre sí, por lo tanto, señales estadísticamente "dependientes". Sepp Hochreiter y Jürgen Schmidhuber mostraron cómo obtener ICA no lineal o separación de fuentes como subproducto de la regularización (1999). Su método no requiere un conocimiento a priori sobre el número de fuentes independientes.

¿Qué es?

El análisis de componentes independientes (ICA) es una técnica estadística y computacional para revelar factores ocultos que subyacen a conjuntos de variables, medidas o señales aleatorias.

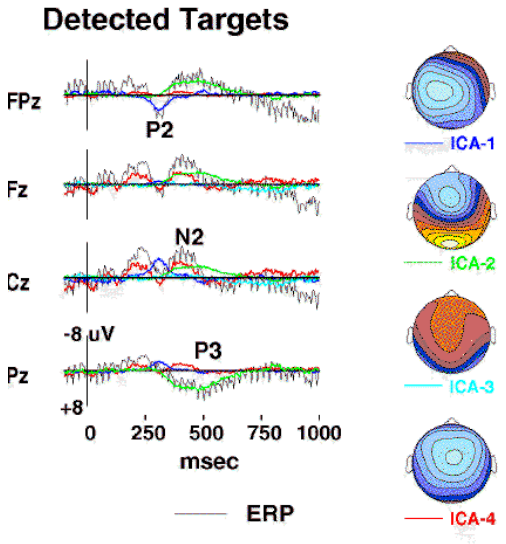
ICA define un modelo generativo para los datos multivariados observados, que normalmente se proporciona como una gran base de datos de muestras. En el modelo, se supone que las variables de datos son mezclas lineales de algunas variables latentes desconocidas, y el sistema de mezcla también se desconoce. Las variables latentes se suponen no gaussianas y mutuamente independientes, y se denominan componentes independientes de los datos observados. ICA puede encontrar estos componentes independientes, también llamados fuentes o factores.





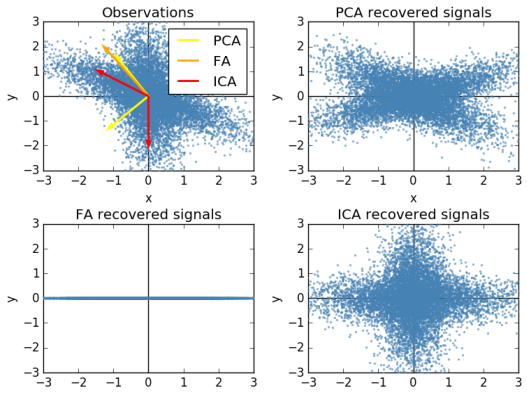
¿Cuál es su objetivo?

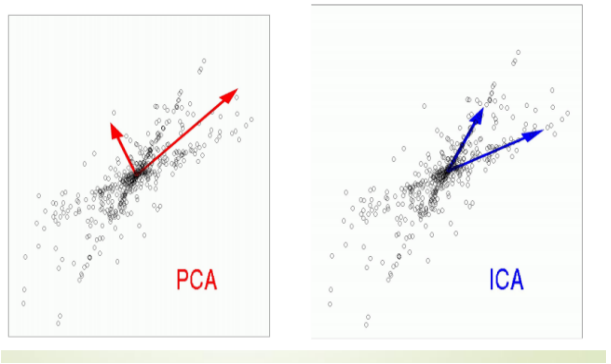
ICA es una técnica emergente para la reducción de dimensión en diagnóstico de fallos, la cual tiene como objetivo encontrar, a partir de una transformación lineal, una representación de un conjunto de variables donde se minimice la dependencia estadística entre las nuevas variables que la forman, llamadas componentes independientes. La independencia estadística entre un conjunto de variables implica que la función de densidad de probabilidad (fdp) conjunta de todas estas es igual al producto de las funciones de densidad de probabilidad marginal de cada una.



¿Con qué Algoritmos se relaciona?

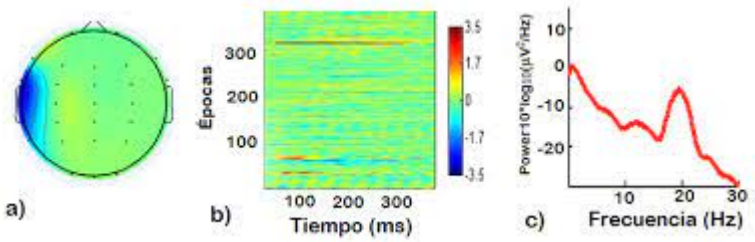
El ICA se relaciona superficialmente con el análisis de componentes principales y el análisis factorial. Sin embargo, ICA es una técnica mucho más poderosa, capaz de encontrar los factores o fuentes subyacentes cuando estos métodos clásicos fallan por completo.





¿De donde provienen los datos analizados por ICA?

Los datos analizados por ICA podrían provenir de muchos tipos diferentes de campos de aplicación, incluidas imágenes digitales, bases de datos de documentos, indicadores económicos y mediciones psicométricas. En muchos casos, las medidas se dan como un conjunto de señales paralelas o series de tiempo; el término separación ciega de la fuente se utiliza para caracterizar este problema. Ejemplos típicos son mezclas de señales de voz simultáneas que han sido captadas por varios micrófonos, ondas cerebrales registradas por múltiples sensores, señales de radio interferentes que llegan a un teléfono móvil o series de tiempo paralelas obtenidas de algún proceso industrial.



Algoritmos que estiman los componentes independientes:

Numerosos algoritmos se han propuesto para estimar las componentes independientes. Sin embargo, uno de ellos es el más reportado por la literatura científica aplicado al diagnóstico de fallos debido a su sencilla implementación y eficiencia computacional, este es conocido como FastICA.

- **Configuración del algoritmo FastICA:**
  El algoritmo FastICA se basa en estimar las componentes independientes a partir del denominado modelo ICA libre de ruido y que se presenta a continuación:

Definición 1.El modelo ICA de un conjunto de n variables aleatorias x = (x1, x2, ..., xn)T , consiste en:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \qquad (1)$$

donde s = (s1, s2, . . . , sn)T es un conjunto de variables aleatorias estadísticamente independientes y A es una matriz cuadrada, llamada matriz de mezcla.

Las dos versiones del algoritmo más aplicadas son FastICA basado en maximización de la no gaussianidad y FastICA basado en estimación de la máxima probabilidad de independencia.22 En el campo del diagnóstico de fallos el primero ha conseguido gran aceptación y se ha aplicado en la mayoría de los trabajos presentados que utilizan esta técnica.9, 14 Este se basa en el Teorema Central del Límite, de forma que si dado un conjunto de variables se busca una combinación lineal de ellas que tenga una función de densidad de probabilidad (fdp) máximamente no gaussiana, se puede encontrar una de las componentes independientes.

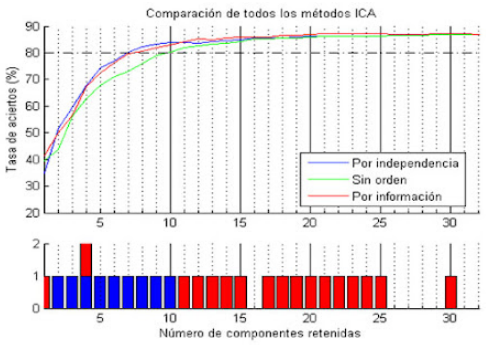


Figura 3. Desempeños del clasificador MAP con FastICA basado en Entropia Negativa con G1

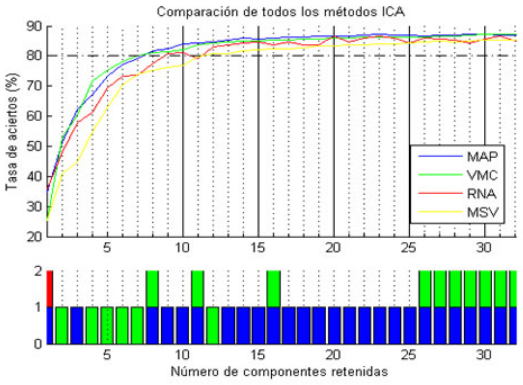
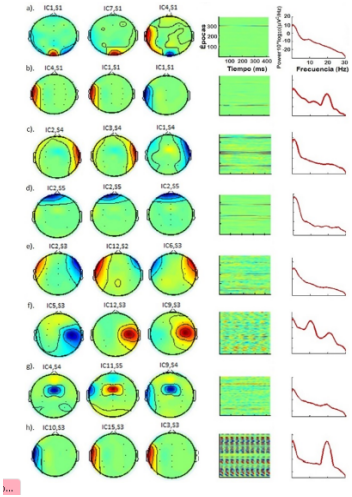


Figura 4. Comparación de los resultados de los clasificadores

- **Infomax.**
  Este algoritmo encuentra la matriz W usando como criterio la minimización de la información mutua entre las fuentes estimadas, con lo que la entropía negativa conjunta se maximiza. Con este criterio, esta implementación hace posible el descomponer señales x en fuentes con distribuciones de probabilidad sub y súper-Gaussianas

- **SOBI. (Second Order Blind Separation)**
  Aunque éste no es estrictamente un algoritmo para el ACI, es una técnica de BBS muy utilizada, por eso se incluyó en este trabajo. Es una implementación que utiliza la ausencia de correlación temporal y espacial entre las fuentes como criterio para definir su independencia, por lo que explota la estructura temporal de las señales para calcular W. Para ello, el algoritmo trabaja sobre una pila de matrices de corrimiento (que construye a partir de las mezclas) y las diagonaliza simultáneamente mediante una matriz de transformación que resulta ser A, la matriz de mezcla e inversa de W.

Artefactos identificados en los registros de EEG utilizando tres algoritmos de preprocesamiento: FastICA, Infomax y SOBI. De izquierda a derecha, mapas topográficos con FastICA, Infomax y SOBI, actividad de las épocas respecto al tiempo y espectro de potencia. (a)-(c) componentes corticales, (d) y (e) artefactos oculares, (f) artefacto muscular, (g) electrodo con alta impedancia y (h) artefacto de origen desconocido. El texto sobre el mapa topográfico indica el número de CI presentado y el sujeto del que fue extraído.

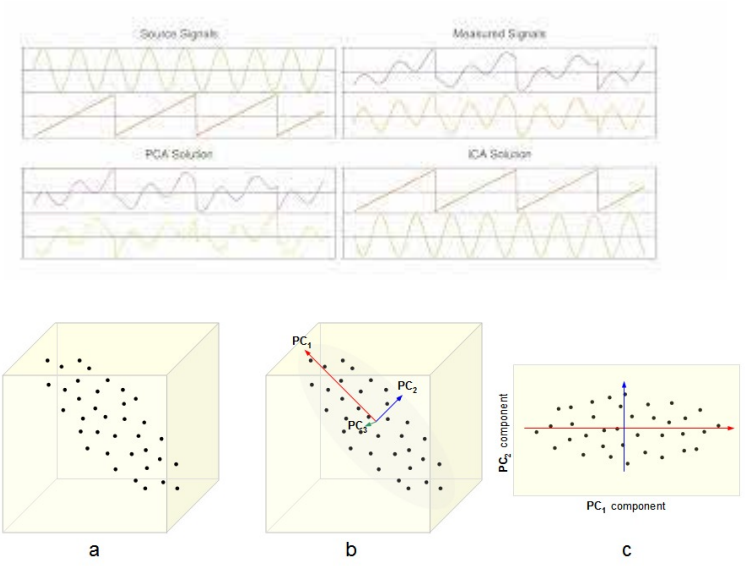


Estrategias para la selección de las componentes independientes al aplicar ICA:
Al contrario de otras técnicas de extracción de características como PCA, que brinda las componentes principales ordenadas de acuerdo al grado de variabilidad que capturan de los vectores de mediciones originales, las variantes del algoritmo FastICA no estiman las componentes independientes siguiendo algún criterio para su disposición. Sin embargo, posterior a la estimación existen varias estrategias a seguir para ordenar las componentes independientes y seleccionar así un número menor de estas para realizar la tarea de clasificación.

Una estrategia muy popular, es privilegiar la selección de las componentes independientes que más variabilidad capturen de las variables originales. Este criterio resulta semejante al que sigue PCA y se denominará como ordenamiento por información (Info).

Otra estrategia puede ser, según los autores del algoritmo FastICA, seleccionar las componentes máximamente no gaussianas de acuerdo al nivel de no gaussianidad que sigan sus distribuciones de probabilidad. Para ello pueden utilizarse diversos estadísticos, pero para ser consecuentes con la parametrización de FastICA se tomará como medida de la no gaussianidad a la Entropía Negativa y se denominará esta forma de ordenamiento por independencia (Ind).

Por último se realizará la selección de las componentes independientes en el mismo orden en que son estimadas de modo natural por el algoritmo FastICA, pues se desea conocer si se consiguen mejoras significativas al reordenarlas siguiendo los criterios mencionados anteriormente.

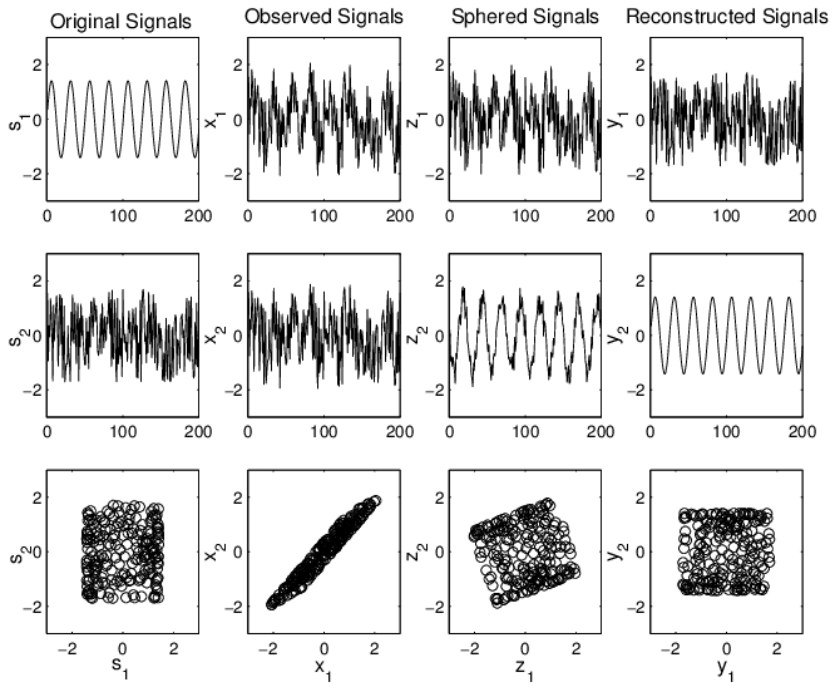


Características

El método ICA tiene múltiples características pero las más destacadas son:

- El número de entradas es igual al número de salidas
- Asume que los componentes independientes son estadísticamente independientes
- Asume que los componentes independientes no son gaussianos (medida de no gaussianidad)
- Las entradas deben ser valores autoescalados (restar cada columna por su media y dividir por su desviación estándar)

Se diferencia del método PCA por su diferencia de enfoque. El PCA se usa para comprimir información (reducción de dimensionalidad) mientras que el ICA se encarga de separar la información.



Metodologías

Análisis de componentes independientes (ICA)

Mediante el cálculo de la estimación de la matriz de mezcla busca minimizar la dependencia estadística entre componentes de las señales originales.

Para ello, es necesario disponer, al menos, del mismo número de mezclas que de fuentes y que, como mucho, solamente una de las fuentes presente una distribución gaussiana.

1. Datos de entrada: Se tendrá los datos de entrada, imágenes, sonidos
2. Formulación del problema: Plantar el problema
3. Hipótesis de aplicación: Si bien la existencia de una única solución, no lo es desde un estricto planteamiento matemático, sí que lo es desde el punto de vista de la independencia de las señales extraídas. En dicho contexto existencia de solución pasa por establecer una serie de condiciones sobre las hipótesis de aplicación que garanticen la separabilidad de las fuentes.
4. Separabilidad de la mezcla:

5. Blanqueado de los datos de entrada: El preproceso de blanqueo de datos antes del algoritmo del método ICA, ayuda a garantizar la convergencia de este y también facilita algunas condiciones para la construcción del algoritmo.

El proceso de blanqueado se lleva a cabo mediante la eliminación de la componente continua de las muestras

El primer paso para el blanqueo consiste en centrar los datos del vector de mezclas x, para esto, restamos la media de cada una de las componentes observadas. Después de esto y a partir de la matriz de covarianza de los datos observados, eliminamos la correlación entre cada una de las señales observadas.

6. Reducir la gaussianidad: la medida de Gaussianidad está relacionada con la independencia estadística entre las variables, entre mayor sea la Gaussianidad, más información comparten las variables y son menos independientes entre sí.

Lo que buscamos es la independencia de las variables.

Caso: Efecto fiesta de Cocktail

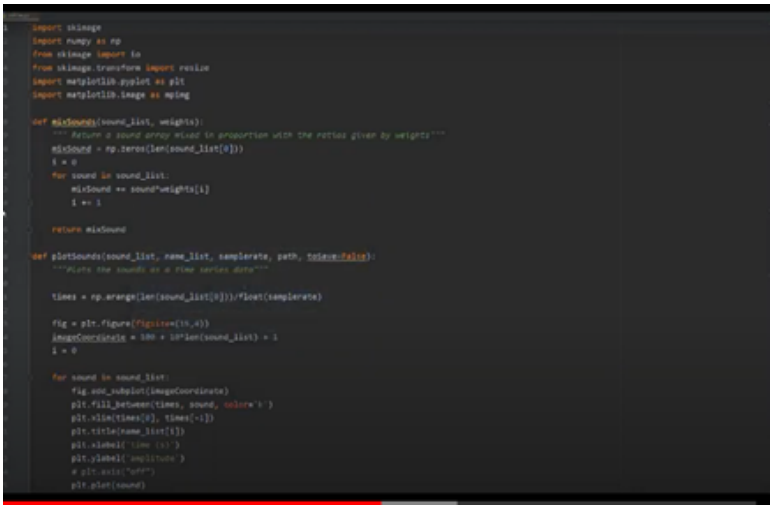
Varias personas en una fiesta, cada una está hablando por lo cual emiten diferentes tipos de ruido, se buscará quien esta hablando y quien no, captar el sonido que deseamos

Nota: personas, variables ocultas

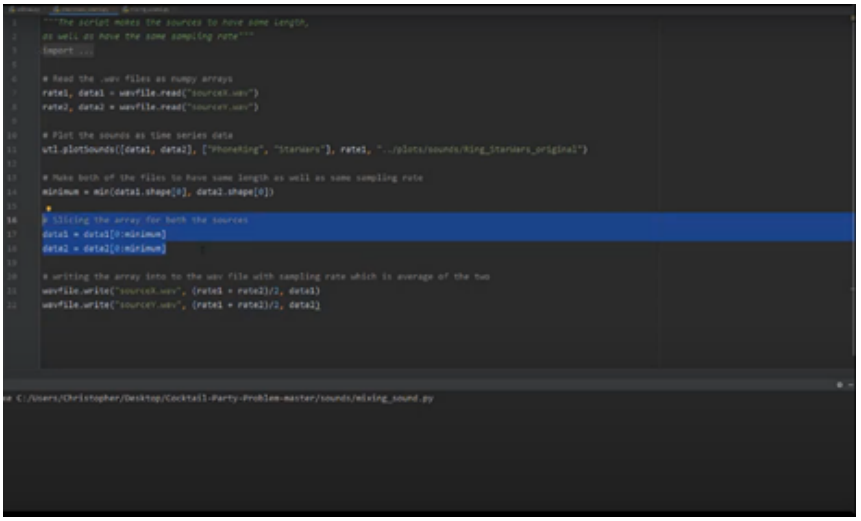
Micrófonos: observadores, los que captan el ruido que emiten cada persona

Cada micrófono captará o grabará todo a su alrededor, de diferente manera, el primer micrófono grabará lo que dice cada una de las personas de manera ligeramente diferente, variara el volumen, retraso del sonido, la intensidad, debido a factores aleatorios como la distancia, el ruido etc.

Se mezclarán los sonidos y se plotearán (mostrarán), para simular la reunión



Se pre-procesaran los datos, se segmentarán las matrices, y se guardarán los archivos pre-procesados



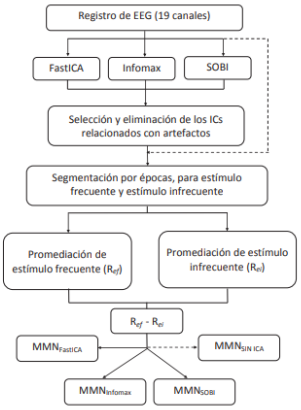
Una Vez que los datos estén pre-procesados, mediante el ICA se leerán los archivos, escalará las variables mezcladas, para luego crear una matriz de las señales y aplicar un blanqueamiento.

Encontrara los componentes individuales uno por uno y aplicar el FaztICA

En el Uso del Análisis por Componentes Independientes en la extracción de artefactos de la respuesta Mismatch Negativity (Ver más) se evidencia que la metodología utilizada fue:

El procedimiento seguido en este trabajo se ha divido en 6 etapas:

1. Descripción de las características de los participantes
2. Registro continuo del EEG durante estimulación auditiva presentada en un paradigma oddball pasivo
3. Obtención de los CIs mediante la aplicación de tres implementaciones del BSS al banco de datos
4. Selección y eliminación sistemática de los CIs correspondientes a artefactos en cada implementación
5. Segmentación en épocas, y cálculo de la respuesta MMN correspondiente
6. Comparación estadística de las respuestas MMN obtenidas al preprocesar usando el BSS versus el MMN obtenido sin preprocesar



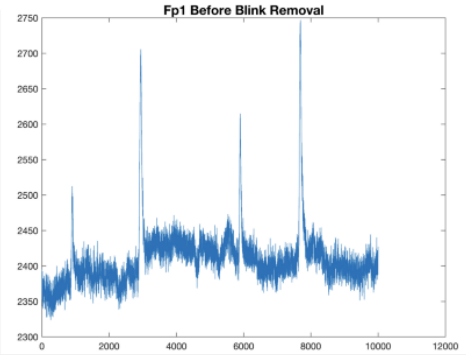
Ejemplo: eliminación de parpadeo de EEG (Tomado de: Aquí)

En este ejemplo, usaré ICA para eliminar artefactos de parpadeo de los datos de EEG, el código está disponible en el repositorio de GitHub .

La electroencefalografía (EEG) es una técnica que mide la actividad eléctrica resultante del cerebro. Una gran desventaja del EEG es su sensibilidad al movimiento y otros artefactos no cerebrales. Uno de esos artefactos ocurre cuando los participantes parpadean. En la siguiente figura, los artefactos de parpadeo se pueden ver claramente a través de picos en el gráfico de voltaje versus tiempo del electrodo Fp1 (cerca de la parte frontal de la cabeza).

Import Data

```
load('EEG_data')

% plot Fp1 electrode data (sits at front left part of head hear left eye)
% perturbations can be seen visually due to blinks
plot(Data(:,1))
title("Fp1 Before Blink Removal", 'FontSize', 14)
```



Importar datos y graficar el voltaje Fp1 frente al tiempo. Imagen del autor.

Un buen primer paso al utilizar ICA es realizar primero PCA en el conjunto de datos. Hacer esto en Matlab se hace fácilmente con la función pca () . Observaré aquí que es fundamental escalar automáticamente los datos. Esto se hace automáticamente en la función pca () . Además, aquí comenzamos con 64 columnas correspondientes a 64 voltajes de electrodo de EEG medidos a lo largo del tiempo. Después de PCA nos quedan 21 columnas correspondientes a 21 vectores de puntuación, es decir, componentes principales.

PCA

```
% DEFINE NUMBER OF PRINICPLE COMPONENTS TO KEEP
% NOTE: it is typically a good idea to try out different numbers and
% compare their total explained variation
q=21;

% PERFORM PCA
% NOTE: it is important to normalize data! (i.e. substract mean of each
% column and divide by standard deviation. MATLAB's pca function does this
% automatically :p
[coeff,Data_PCA,latent,tsquared,explained,mu] = pca(Data, 'NumComponents', q);

% compute and display explained variation
disp(strcat("Top ", string(q), " principle components explain ", ...
    string(sum(explained(1:q))), " of variation"))
```

```
Top 21 principle components explain 99.5134 of variation
```

Código para aplicar PCA al conjunto de datos. Imagen del autor.

A continuación, podemos entrenar un modelo ICA y aplicarlo a la matriz de puntuación de PCA.

ICA

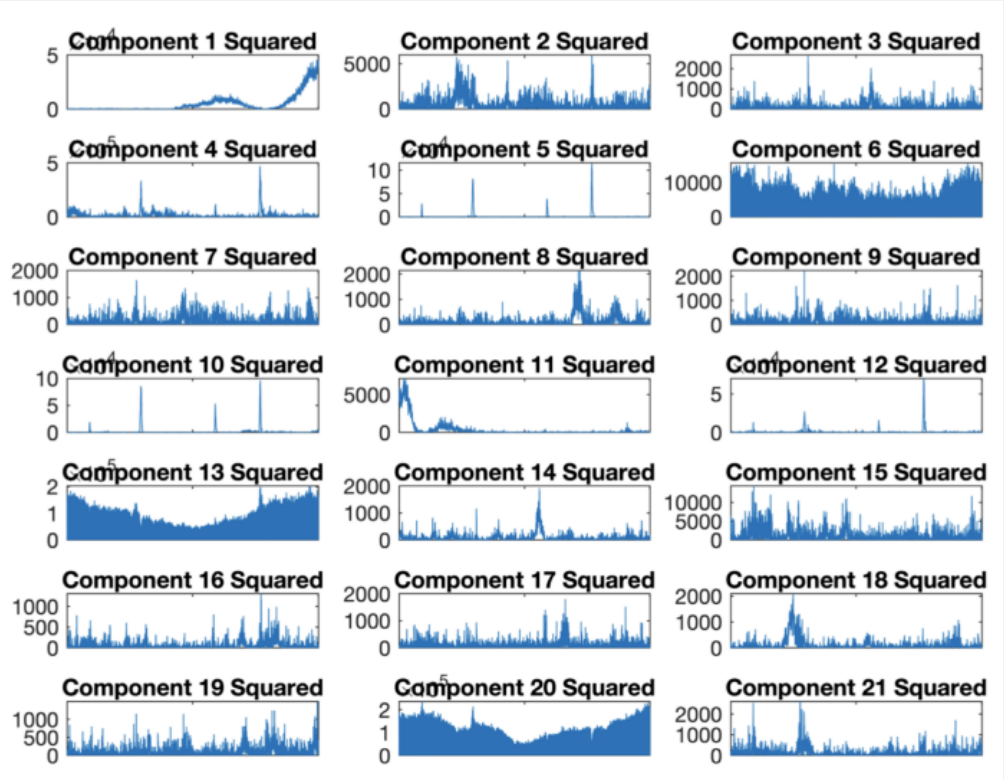
```
% compute independent components from principle components
% train ICA model
Mdl = rica(Data_PCA, q);

% apply ica
Data_ICA = transform(Mdl, Data_PCA);
```

Plot Components

```
% define number of plots per column of figure
plotsPerCol = 7;

% plot components
figure(2)
for i = 1:q

    subplot(plotsPerCol,ceil(q/plotsPerCol),i)
    plot(Data_ICA(:,i).^2)
    title(strcat("Component ", string(i), " Squared"))
    ax = gca;
    ax.XTickLabel = {};

end
```

Código para aplicar ICA a componentes principales. Imagen del autor.

Trazando los componentes independientes, podemos inspeccionar cuáles corresponden a artefactos parpadeantes.



Gráficas de 21 componentes independientes al cuadrado. Imagen del autor.

Utilizo una heurística perezosa para seleccionar componentes independientes que representan información de parpadeo. Es decir, la selección de componentes cuyo cuadrado tiene 4 picos prominentes. Los componentes restantes se pueden usar para reconstruir el conjunto de datos original sin información de estos componentes de parpadeo.

Remove Embedded Blink Information

```
% use heuristic to pick component corresponding to blink
Components_blink = pickBlinkComponents(Data_ICA);
disp("Blink component(s):")
```

```
Blink component(s):
```

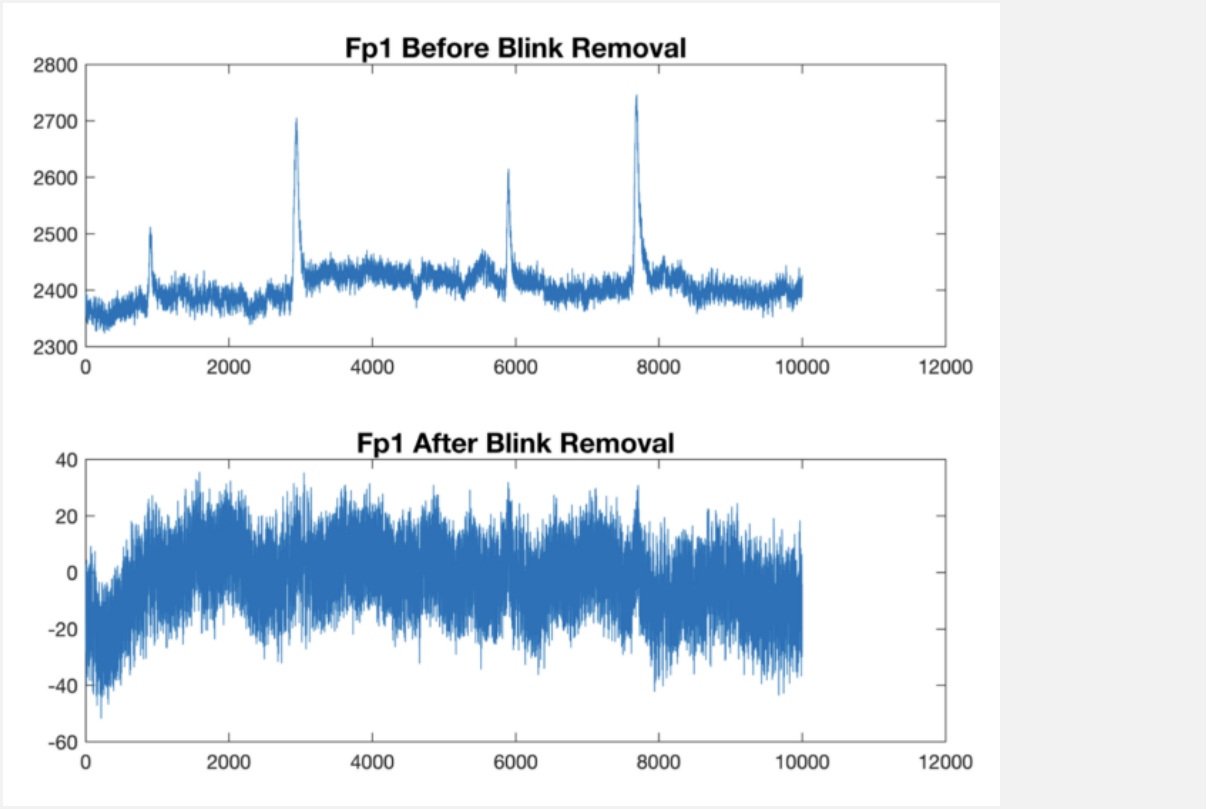
```
disp(Components_blink)
```

```
     3    20
```

```
% zero all columns corresponsing to blink components
Data_ICA_noBlinks = Data_ICA;
Data_ICA_noBlinks(:,Components_blink) = ...
    zeros(length(Data_ICA), length(Components_blink));

% perform inverse ica transform
Data_PCA_noBlinks = Data_ICA_noBlinks*Mdl.TransformWeights;

% perform inverse pca transform
Data_noBlinks = Data_PCA_noBlinks*coeff';

% plot Fp1 electrode before and after
figure(3)
subplot(2,1,1)
```

Código para seleccionar componentes independientes de parpadeo y reconstruir datos de EEG. Imagen del autor.

Finalmente, graficamos el voltaje original y resultante a lo largo del tiempo para el electrodo Fp1.



Señal Fp1 antes y después de la eliminación de parpadeo.

Aplicación del ICA

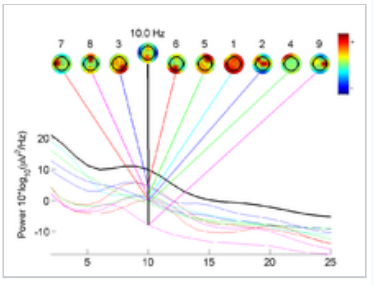
Al realizar una investigación pudimos obtener que el ICA ha tenido variadas aplicaciones y unas de las que identificamos fueron:

1. Análisis financiero (predecir los precios del mercado de valores)
2. Neurociencia (clasificación de picos neuronales)
3. Análisis de Componentes Principales e Independientes aplicados a Reducción de Ruido en Señales Electrocardiográficas. Ingrese Aquí
4. Aplicaciones del Análisis de Componentes Independientes al procesamiento de registros de movimientos oculares sacádicos. Ingrese Aquí
5. Análisis de componentes independientes aplicado al estudio de la actividad cerebral. Ingrese Aquí
6. Análisis de Componentes Independientes Aplicado a Series Financieras - Ingrese Aquí
7. Análisis de Componentes Independientes en Separación de Fuentes de Ruido de Tráfico en Vías Interurbanas - Ingrese Aquí
8. ICA Aplicado a la Extracción de Características en Imágenes. Ingrese Aquí

ICA se puede ampliar para analizar señales no físicas. Por ejemplo, ICA se ha aplicado para descubrir temas de discusión en una bolsa de archivos de listas de noticias.

Algunas aplicaciones ICA se enumeran a continuación:

- Imágenes ópticas de neuronas
- Clasificación de picos neuronales
- Reconocimiento facial
- Modelado de campos receptivos de neuronas visuales primarias
- Predecir los precios del mercado de valores
- Comunicaciones por telefonía móvil
- Detección basada en el color de la madurez de los tomates
- Eliminar artefactos, como parpadeos, de los datos del EEG .
- Análisis de los cambios en la expresión génica a lo largo del tiempo en experimentos de secuenciación de ARN de una sola célula .
- Estudios de la red en estado de reposo del cerebro.
- Astronomía y cosmología



Análisis de componentes independientes en EEGLAB

Aplicación del análisis de componentes independientes (ICA) en la interpretación de señales electroencefalográficas en epilepsia

El análisis de componentes independientes (ICA) es una técnica novedosa en el estudio de la señal electroencefalográfica. Los estudios previos son escasos y parciales. Los objetivos de este trabajo han sido determinar de forma cuantificada la validez de ICA en la eliminación de artefactos, comparar ICA y filtros digitales en la mejoría obtenida en la visualización del inicio de crisis, estudiar la fisiopatología de descargas epileptiformes intercríticas y crisis de distintos tipos de epilepsia y comprobar la capacidad de ICA para diferenciar entre epilepsias uni y multifocal es. Para ello se aplicó ICA a muestras de eeg intercrítico (muestras artificiales y muestras reales) Y crítico, y se reconstruyeron con los componentes de interés. Los resultados fueron valorados mediante técnicas de análisis frecuencial, proyección de topografías y cálculo de correlaciones, los resultados indican que ICA elimina de forma selectiva los artefactos en registros de EEG. Su eficacia depende de la amplitud y morfología del artefacto, del número de componentes calculados y el montaje utilizado. ICA obtiene mejores resultados que los filtros digitales para mejorar la visualización del inicio de las crisis epilépticas; la combinación de ambas técnicas fue la mejor opción. ICA descompone las grafoelementos intercríticos y las crisis en componentes con una secuencia temporal y topográfica. La separación de las descargas intercríticas varía dependiendo del tipo de epilepsia mientras que la de las crisis depende del tipo de patrón electroencefalográfico. ICA diferencia entre epilepsias uni y multifocales. Estos datos indican que ICA es una herramienta muy útil para el estudio de distintos aspectos del EEG.

REFERENCIAS

https://www.cs.helsinki.fi/u/ahyvarin/whatisica.shtml
http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1815-59282014000200007#e1
http://www.scielo.org.mx/pdf/rmib/v38n2/2395-9126-rmib-38-02-00420.pdf
https://en.wikipedia.org/wiki/Independent_component_analysis
https://dialnet.unirioja.es/servlet/tesis?codigo=276153
https://towardsdatascience.com/independent-component-analysis-ica-a3eba0ccec35
http://www.scielo.org.mx/scielo.php?pid=S0188-95322017000200420&script=sci_arttext
http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1815-59282014000200007
https://repositorio.unal.edu.co/bitstream/handle/unal/52081/01830451.2014.pdf?sequence=1&isAllowed=y
https://www.youtube.com/watch?v=l6CJAXcjcWE