

Neuroproject

Estadística Aplicada III



Aldo Emmanuel Carmona Jiménez 196950
Diego Arellano Zamudio 198002
Mateo De La Roche 190748
Víctor Contreras 198602

Introducción

Este trabajo final para el curso de Estadística Aplicada III se enfoca en explorar y comparar dos métodos de reducción de dimensionalidad prominentes: el Análisis de Componentes Principales (PCA) y el t-distributed Stochastic Neighbor Embedding (t-SNE). Estos métodos serán aplicados y evaluados a través de dos conjuntos de datos de señales cerebrales. El primer conjunto consiste en datos recogidos mientras los sujetos realizaban una tarea específica con el dedo, lo cual implica una actividad neuronal focalizada. El segundo conjunto comprende señales cerebrales captadas mientras los sujetos observaban imágenes de animales específicos (perros, gatos, conejos) y un grupo de control constituido por imágenes aleatorias. El objetivo principal de este estudio es describir y discernir las diferencias en la efectividad del PCA y el t-SNE para visualizar y posiblemente segmentar las señales cerebrales en contextos distintos: uno relacionado con una tarea motora simple y otro vinculado a la respuesta neuronal ante estímulos visuales variados. El PCA, conocido por su capacidad para reducir la dimensionalidad manteniendo la mayor varianza posible, se contrastará con el t-SNE, que es altamente efectivo en la modelación de la estructura de los datos en un espacio de baja dimensión, resaltando agrupaciones y patrones no lineales.

Al final, se espera que los resultados obtenidos aporten a la mejora de las estrategias de análisis en estudios neurocientíficos, facilitando así la interpretación más precisa de las complejas actividades cerebrales.

Datos

El primer conjunto de datos analizado fue recopilado por Kai Joshua Miller de la universidad de Stanford, en este trabajo se utilizará una porción de éstos que corresponde a 3 sesiones de 3 sujetos. Miller describe la forma en que recopilaron los datos de la siguiente forma: En la tarea de movimiento de los dedos, los sujetos recibieron indicaciones con una palabra mostrada en un monitor junto a la cama que indicaba qué dedo mover durante pruebas de movimiento de 2 segundos. El sujeto realiza movimientos a su propio ritmo en respuesta a cada una de estas señales y, por lo general, movió cada dedo de 2 a 5 veces durante cada prueba, pero algunas pruebas incluyeron muchos más movimientos. Las variables de este conjunto de datos son las siguientes:

1. 'V': Esta variable representa los datos de voltaje continuo, organizados en una matriz de tiempo por canales. Cada fila corresponde a un punto en el tiempo y cada columna a un canal específico. Este tipo de datos es fundamental para el análisis de señales electroencefalográficas (EEG) o cualquier otro tipo de registro eléctrico cerebral, permitiendo estudiar la actividad eléctrica del cerebro en respuesta a diversos estímulos o acciones.
2. 'locs': Contiene las coordenadas tridimensionales de los electrodos utilizados para registrar las señales. Estas coordenadas son esenciales para identificar la ubicación física de los electrodos en la cabeza, lo que puede correlacionarse con diferentes regiones cerebrales implicadas en los movimientos de los dedos.
3. 'dg': Esta variable representa las flexiones de los dedos, estructurada en una matriz de tiempo por número de dedos. Los datos fueron preprocesados para remover una línea de base móvil, lo que sugiere que se han ajustado para eliminar variaciones lentas que podrían distorsionar la interpretación de los datos de flexión rápida de los dedos.
4. 'srate': Indica la tasa de muestreo de los datos, la cual es de 1000 Hz. Esto significa que los datos fueron muestreados 1000 veces por segundo, proporcionando una resolución temporal alta que es crucial para capturar la rápida dinámica de la actividad cerebral y los movimientos de los dedos.

5. 't_on' y 't_off': Estas variables marcan el inicio y el fin de los estímulos respectivamente. 't_on' es el tiempo en que se presenta el estímulo para flexionar un dedo, y 't_off' generalmente ocurre 2000 ms después de 't_on', señalando el fin de este estímulo. Estos marcadores son útiles para aislar segmentos de datos correspondientes a períodos específicos de interés alrededor de los estímulos.
6. 'stim_id': Identidad del estímulo que indica cuál dedo debe flexionarse. Esta información es crucial para estudios que investigan cómo diferentes dedos son controlados por el cerebro y cómo esta información es procesada a nivel neuronal.
7. 'stimtext': Texto del estímulo mostrado en la pantalla, probablemente instrucciones directas para el sujeto sobre qué dedo debe mover. Este detalle ayuda a correlacionar los datos de voltaje y movimiento con la tarea específica que estaba siendo realizada en ese momento.

Al hacer el análisis exploratorio de los datos se pudo ver que el número de electrodos para los 3 sujetos son diferentes.

Después se aplicó un filtro de Butterworth cuyo objetivo es tener una respuesta en frecuencia lo más plana posible en la banda de paso, para lograrlo, se aplicaron dos filtros a las señales. El primer filtro elimina las frecuencias por debajo de 50 Hz, que suelen incluir ruido de baja frecuencia como el movimiento lento y el segundo filtro tiene una frecuencia de corte de 3 Hz, este filtro suaviza la señal de potencia eliminando variaciones de alta frecuencia y resaltando las tendencias de baja frecuencia.

El segundo conjunto de datos fue recopilado por el Neurolab del Instituto Tecnológico Autónomo de México (ITAM). El método por el que fueron recopilados fue el siguiente: cada sujeto veía imágenes en una pantalla de perros, gatos, conejos y de control (imágenes aleatorias), cada imagen duraba aproximadamente un segundo en pantalla. La frecuencia del muestreo fue de 125 Hz y cada observación es de 123 Hz, poco menos de un minuto. Las variables de este conjunto de datos son las siguientes:

1. 'egg_signal': Señales del electroencefalograma (EEG) que mide la actividad eléctrica del cerebro.

2. 'class': Imagen que se le fue enseñada al sujeto.
3. 'subject': Nombre de la persona.
4. 'age': Edad de la persona, entre 22 y 24.
5. 'sex': Sexo de la persona, la gran mayoría son hombres.
6. 'has_cat': Variable binaria que indica si la persona tiene un gato como mascota.
7. 'has_dog': Variable binaria que indica si la persona tiene un perro como mascota.
8. 'has_rabbit': Variable binaria que indica si la persona tiene un conejo como mascota.

Debe de considerarse que la gran mayoría de los participantes en este conjunto de datos son hombres entre 22 y 24 años y hay una persona repetida aunque estaba sujeta a diferentes condiciones. A estas señales no se le aplicó el filtro de Butterworth.

Métodos

Centramos el análisis en verificar la capacidad con la que PCA y t-SNE nos otorgan resultados visuales simples e interpretables. Al tratarse de datos caóticos como lo son los neurotransmisores, esperamos notar una importante diferencia en la capacidad de tratar con ellos según el método.

A pesar de que ambos son métodos utilizados para la reducción de dimensionalidad, cada uno tiene territorios en los que su aplicabilidad resulta mejor que la del otro. Nuestra misión es verificar qué tipo de territorio es la neurociencia, y por lo tanto, qué enfoque utilizar. Para este fin, vamos a conocer más a fondo los supuestos que maneja cada técnica, el propósito de cada uno y la interpretación que debemos tomar de los resultados.

Por un lado, PCA se utiliza principalmente para reducir la dimensionalidad de los datos preservando tanta variabilidad como sea posible. Esto se logra proyectando los datos en un conjunto de nuevas variables (componentes principales) que son combinaciones lineales de las variables originales. Por otro lado, t-SNE se utiliza principalmente para la visualización de datos de alta dimensionalidad en dos o tres dimensiones, manteniendo las relaciones locales entre los puntos de datos.

PCA es un método lineal. Encuentra las direcciones (componentes principales) en las que los datos varían más y proyecta los datos en estas direcciones. t-SNE es un método no lineal. Se centra en preservar la estructura local de los datos, es decir, puntos cercanos en el espacio original deberían permanecer cercanos en el espacio reducido, y viceversa.

Los componentes principales son ortogonales entre sí, lo que significa que no están correlacionados. Los primeros componentes capturan la mayor parte de la variabilidad presente en los datos. t-SNE no produce una transformación lineal de los datos. En cambio, se enfoca en la proximidad de los puntos, haciendo que sea difícil interpretar la relación directa con las variables originales.

PCA es computacionalmente eficiente y puede manejar grandes conjuntos de datos, ya que su complejidad computacional es relativamente baja. t-SNE es computacionalmente más intensivo que PCA, lo que puede limitar su aplicación en conjuntos de datos muy grandes.

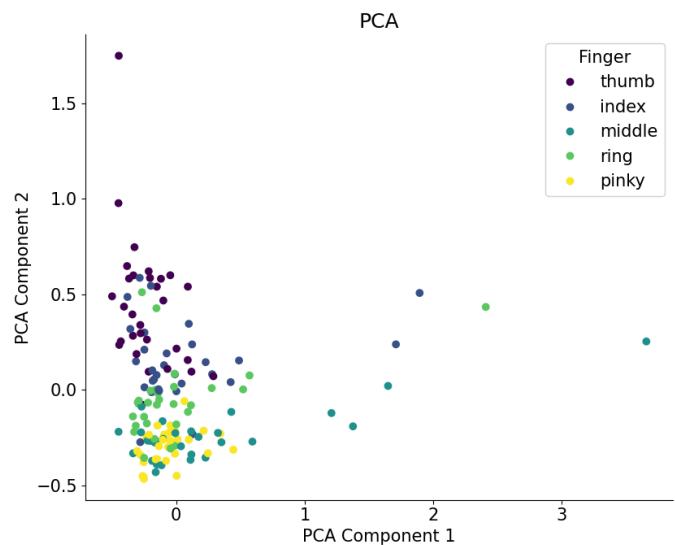
La salida de PCA son nuevas características (componentes principales) que son combinaciones lineales de las características originales. La salida de t-SNE son coordenadas en un espacio de menor dimensión (generalmente 2D o 3D) que pueden ser usadas para visualización, pero no son combinaciones lineales de las características originales.

Resultados

Tenemos dos métodos por contrastar y dos datasets con los cuales podemos contrastarlos. Este apartado estará dividido en 4 secciones que corresponden a los resultados visuales que arrojan cada uno de los dos métodos a cada uno de los conjuntos de datos estudiados.

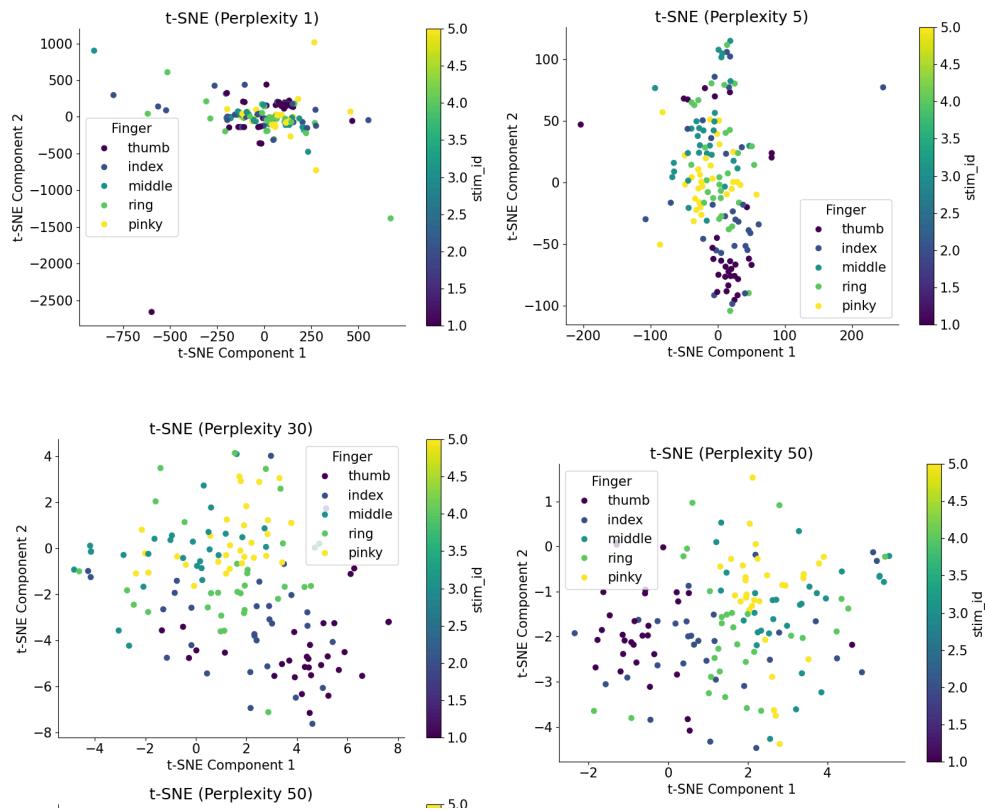
Cabe decir que t-SNE tiene un parámetro que resultó importante para este estudio: perplexity. La perplexity puede interpretarse como una medida de la cantidad de vecinos efectivos que cada punto considera en el espacio de alta dimensionalidad. Es una manera de ajustar la "escala" a la cual t-SNE observa la proximidad entre puntos de datos. En los resultados veremos 4 gráficas de t-SNE que corresponden a distintos valores del parámetro perplexity: 1, 5, 30 y 50.

PCA para las actividades específicas con los dedos



En la gráfica podemos observar que no existe una agrupación que resalte visualmente. Nuestro interés se enfoca en la distribución de los puntos según el dedo que generó tales métricas. El meñique y en parte el pulgar son los dedos que más se mantienen agrupados entre sí. Los otros dedos tienen algunos datos que salen de sus respectivas agrupaciones.

t-SNE para las actividades específicas con los dedos



Para el t-SNE con perplejidad 1 los puntos están altamente dispersos y no forman agrupaciones distinguibles, esto sugiere que una perplejidad baja es inadecuada para estos datos, ya que no permite que el algoritmo capture ninguna estructura significativa de las relaciones entre puntos.

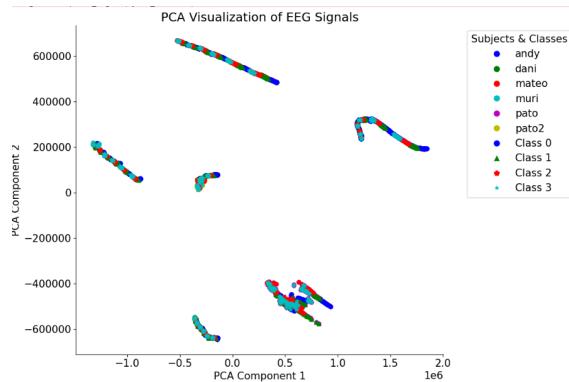
Para el t-SNE con perplejidad 5, los puntos siguen dispersos y sin formar agrupaciones claras, parecen estirarse a lo largo de la dimensión vertical.

El t-SNE con perplejidad 30 es similar al caso anterior, pero con una ligera mejora en la separación de los grupos. Las agrupaciones están un poco más definidas, especialmente para el pulgar y el meñique.

Finalmente, para el t-SNE con perplejidad 50 la distribución de los puntos en el espacio bidimensional muestra cierta agrupación por colores, que representan diferentes dedos. Aunque hay cierta superposición, se pueden observar grupos definidos especialmente para los dedos pulgar, índice y meñique. Los resultados

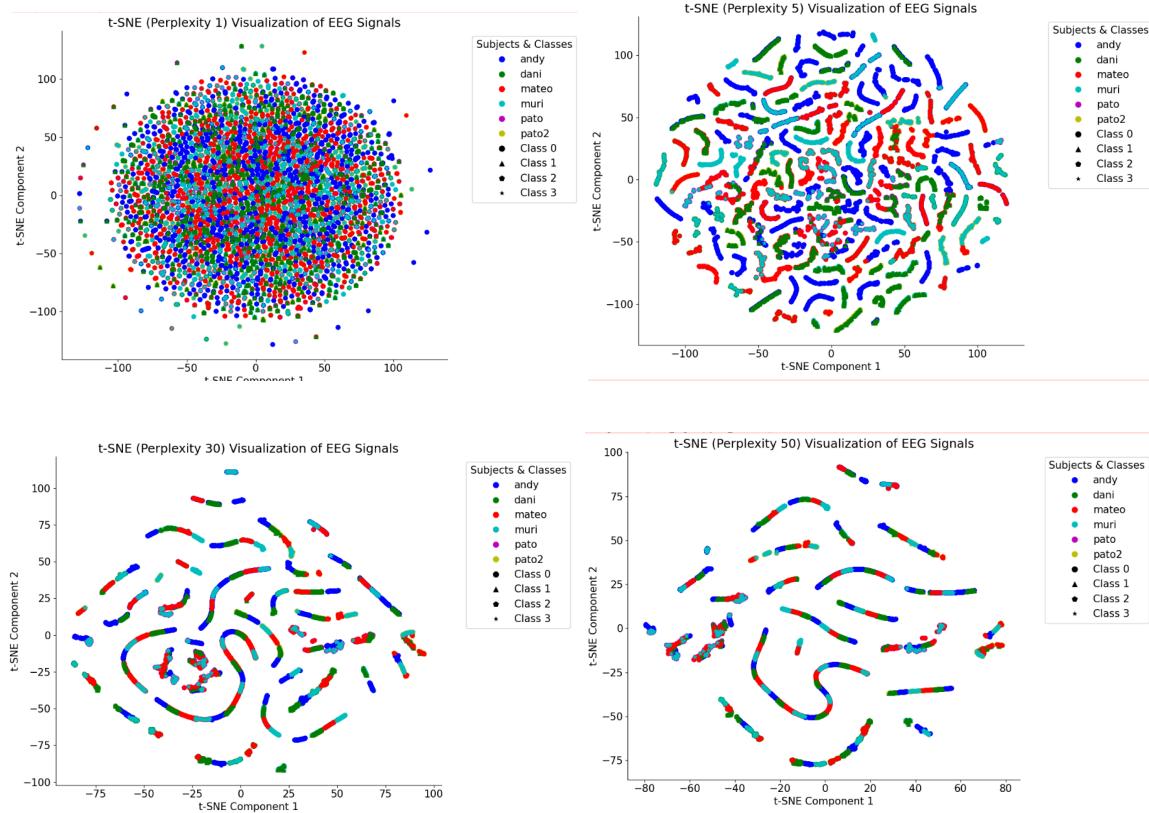
indican que una perplejidad moderada (en torno a 30-50) es más efectiva para este tipo de datos EEG.

PCA para los estímulos visuales



Los datos están distribuidos en varios grupos, indicando que las señales de EEG tienen una estructura clara cuando se proyectan en el espacio de las dos primeras componentes principales. Esto sugiere que PCA es capaz de captar diferencias significativas en las señales de EEG relacionadas con las imágenes mostradas y los sujetos.

t-SNE para los estímulos visuales

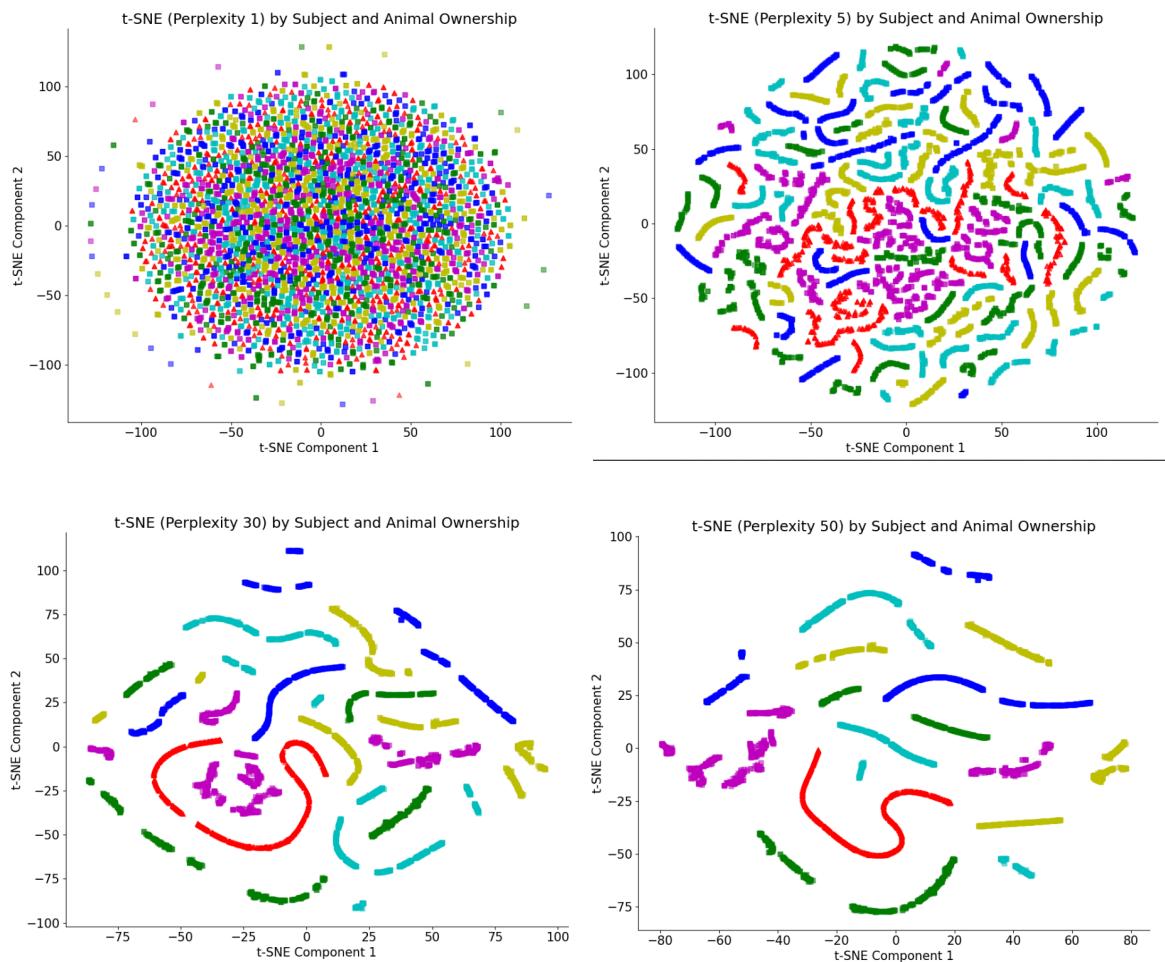


La visualización de los datos EEG utilizando t-SNE con diferentes niveles de perplejidad muestra cómo la elección de este hiper parámetro afecta la agrupación y separación de los datos.

Para el t-sne con perplejidad de 50 la distribución de los puntos muestra varias agrupaciones largas y serpentinas. Los puntos de diferentes sujetos y clases están mezclados en estas formas. Una perplejidad de 50 capta relaciones de mediana escala entre los datos, permitiendo la formación de agrupaciones coherentes pero aún con mezcla de sujetos y clases. Esto sugiere que hay una estructura subyacente en los datos, aunque no suficientemente clara para separar por completo las clases.

Las perplejidades altas (50 y 30) capturan relaciones de mediana escala, resultando en estructuras largas y serpentinas. Adecuadas para ver la estructura global de los datos, aunque no separan claramente las clases.

Las perplejidades bajas (5 y 1) parecen enfatizar relaciones locales, resultando en agrupaciones densas y menos claras. Las perplejidades muy bajas pueden llevar a una pérdida de la estructura general de los datos.



Conclusiones

PCA y t-SNE son herramientas poderosas en el análisis de datos, y cada una tiene ventajas y limitaciones específicas para el análisis de datos de neurotransmisores. PCA se destaca en situaciones donde se requiere una reducción de dimensionalidad rápida y eficiente, preservando la mayor cantidad de variabilidad original de los datos. Este método es ideal para tareas como la simplificación de modelos predictivos, la eliminación de ruido y la compresión de datos. Su naturaleza lineal permite interpretaciones claras de los componentes principales, facilitando el entendimiento de la estructura subyacente de los datos de neurotransmisores.

En contraste, t-SNE se muestra como una opción excelente para la visualización de datos de neurotransmisores debido a la naturaleza caótica y no lineal de estos datos. La capacidad de t-SNE para preservar las relaciones locales entre puntos hace que sea una herramienta invaluable para descubrir patrones y agrupamientos en los datos de neurotransmisores que no son evidentes a simple vista. Esta técnica es especialmente valiosa en el análisis exploratorio de estos datos complejos, donde las interacciones no lineales son comunes y la visualización intuitiva puede revelar estructuras importantes. Sin embargo, su mayor costo computacional y la dificultad para interpretar las coordenadas resultantes limitan su uso a conjuntos de datos de tamaño moderado y propósitos de visualización, que para nuestros fines fue funcional.

La escalabilidad es otra área donde PCA y t-SNE difieren significativamente. PCA, siendo computacionalmente menos intensivo, puede manejar grandes conjuntos de datos de neurotransmisores con relativa facilidad, lo que lo hace adecuado para análisis de datos en tiempo real y aplicaciones que requieren procesamiento rápido. Por otro lado, t-SNE, debido a su complejidad, es más adecuado para análisis más detallados y específicos, donde la calidad de la visualización es crítica y el tiempo de procesamiento no es una limitación tan severa.

En términos de aplicación, PCA es una elección común en análisis de regresión, clasificación y detección de anomalías en datos de neurotransmisores, donde la reducción de dimensionalidad mejora la eficiencia y precisión de los modelos.

t-SNE, sin embargo, se puede utilizar principalmente en la fase exploratoria del análisis de datos de neurotransmisores, ayudando a los analistas a identificar agrupamientos naturales y estructuras intrínsecas en los datos que luego pueden ser investigadas más a fondo utilizando otras técnicas.

Ambos métodos tienen roles complementarios en el análisis de datos de neurotransmisores. La elección entre PCA y t-SNE depende del objetivo específico del análisis, el tamaño y la naturaleza de los datos, y los requisitos de interpretación y visualización del usuario. Al entender estas diferencias, podemos seleccionar la herramienta más adecuada para cada caso particular de investigación y maximizar la información extraída de los datos de neurotransmisores.

Fuentes

- Miller K. J. (2019). *A library of human electrocorticographic data and analyses*. Stanford Libraries. Recuperado de:
<https://exhibits.stanford.edu/data/catalog/zk881ps0522>
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579-2605.
- NeuroLab-ITAM(2023) SynapSee_data