

Reducción de dimensiones.

PCA: Análisis de componentes principales.

Dimiensionalidad

Marca
Costo
Color
No. Pasajeros
Vestiduras
No. Cilindros
Frenos
Transmisión



La maldición de la dimensionalidad.

- ¿Supongamos que tenemos N puntos uniformemente distribuidos en un hypercubo de dimensión D . El hypercubo está centrado en cero, y tiene una longitud $2r$.
 - ¿Cuál es la proporción de puntos que caen dentro de una unidad r de distancia a partir del origen?
 - ¿Cómo varia esa proporción conforme aumenta la dimensión D ?
- Ejercicio: Responde matemáticamente a dichas preguntas, inicia con $D=2$, $D=3$, etc, y calcula el límite cuando $D \rightarrow \infty$.

La maldición de la dimensionalidad.

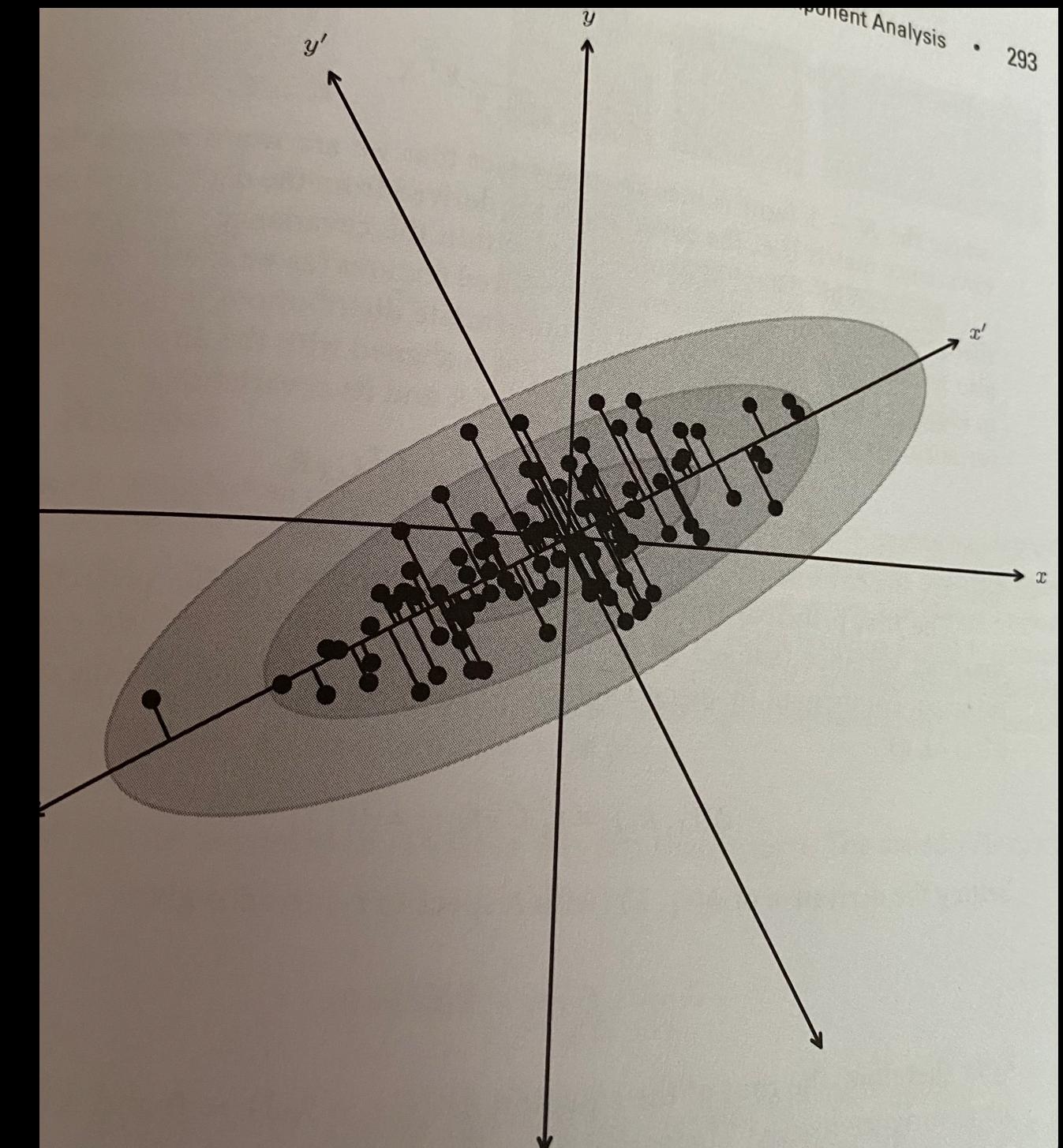
- Si los datos tienen muchas propiedades asociadas, y estas son mas que las observaciones independientes, podemos estar sobre ajustando nuestro modelo.
- Es más difícil encontrar patrones, o clusters en datos con muchas propiedades, cada observación en el conjunto de datos parece igualmente equidistante.
- La probabilidad de encontrar una observación con una combinación particular de las propiedades tiende a cero.
- En casos como estos será útil reducir la dimensionalidad.
- Ej. Espectros de galaxias/cuásares en SDSS (ver ejercicio del 2 de Octubre de 2020: eboss_qso_DR14.ipynb)

Reducción de dimensiones del espacio de variables

- Eliminación de variables. ——> No todas las variables afectan de igual forma a la observable. Ignoramos aquellas que la afectan menos.
- Extracción de variables. Creamos nuevas variables ordenadas respecto a que tan bien predicen la observable. Usamos solo las n-primeras nuevas variables. E.g. PCA
- PCA: Análisis de Componentes Principales
 - Con: Las nuevas variables dejan de ser fácilmente interpretables.
 - Pros: Se reduce el número de variables sin tener que hacer juicios sobre cuál es mas o menos importante; las nuevas variables son independientes; Podemos hacer regresiones o inferencia con las nuevas variables.

PCA: Análisis de componentes principales.

- Objetivo: encontrar combinaciones de variables que sean independientes y que estén ordenadas por importancia o contribución a la observable.
 - Para N observaciones con K variables los datos se ordenan en una matriz de $N*K$ dimensiones. Ej. Flujo por bin de longitud de onda, para N espectros.
 - El objetivo es encontrar los eigenvalores y eigenvectores de esa matriz. Ya que cualquier vector de la matriz original puede construirse a partir de una combinación lineal de los eigenvectores.
- Los eigenvectores coinciden con las direcciones de máxima variación.



PCA

- 1) Organiza los datos en una matriz de N filas y K columnas, donde cada columna corresponde con las variables independientes.
- 2) Centrar y estandarizar la matriz X .
 - Centrar: restar el promedio en cada columna, para que una tenga promedio cero.
 - Estandarizar: Dividir la variable en cada columna por su desviación estándar, esto es que cada columna tendrá desviación estándar 1. Esto se hace solo si la varianza de cada variable no importa, pero puede omitirse si es que es importante tener información de la varianza en cada variable.
- 3) Encontrar las componentes principales R.
 - Opción a: Encontrar los eigenvalores de la matriz de covarianza $C_X = \frac{1}{N-1}X^T X$
 - Ej. Multiplicadores de Lagrange.
 - Opción a: Encontrar los eigenvalores de la matriz de covarianza $M_X = \frac{1}{N-1}XX^T$
 - Ej. Multiplicadores de Lagrange

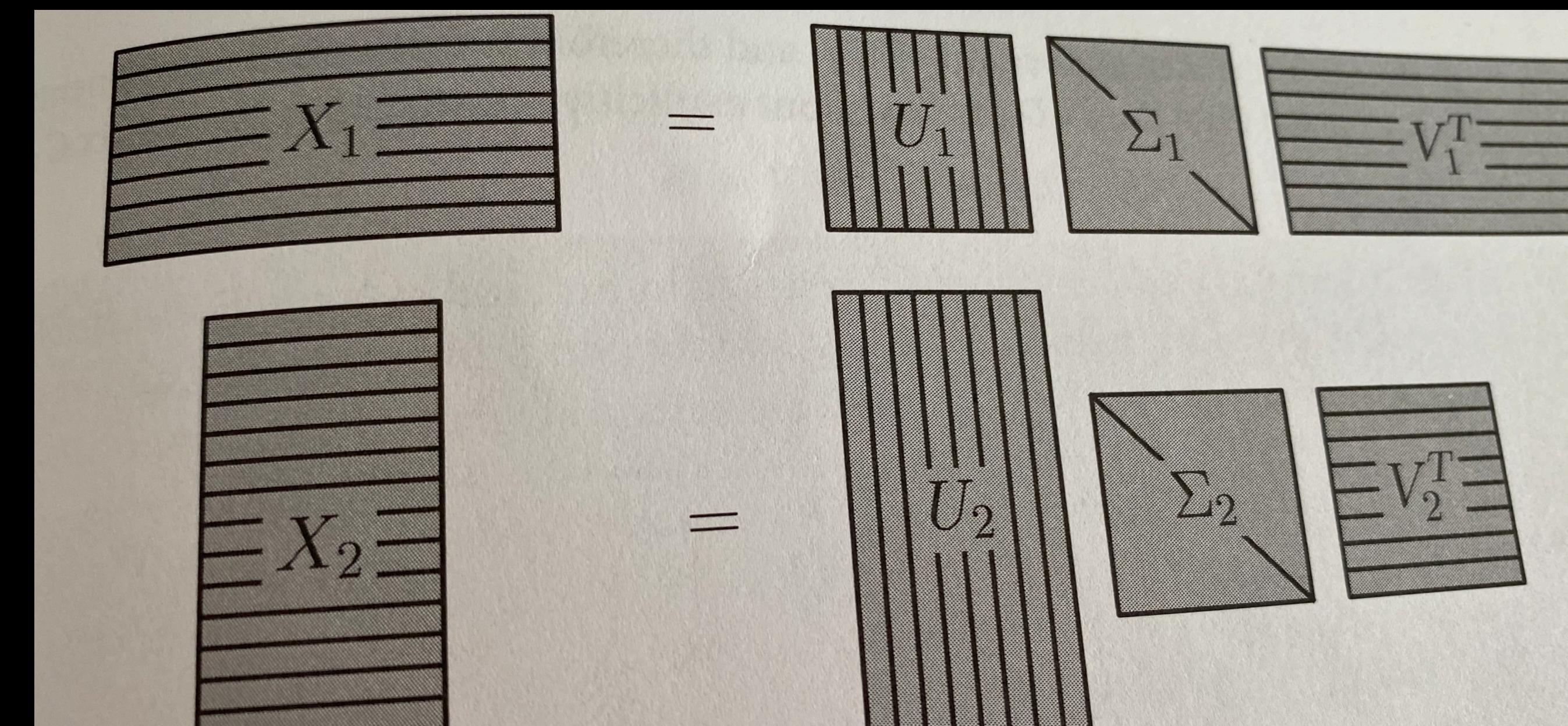
PCA

Opción c) Calcular los eigenvectores y eigenvalores de X, ej. Singular Value Decomposition.

$U\Sigma V^T = \frac{1}{\sqrt{N-1}}X$, U: vectores singulares izquierdos, V vectores singulares derechos, Σ matriz de valores singulares de tamaño RxR con R=min(N,K).

Los vectores singulares derechos V, corresponden con las componentes principales

La forma de U o V depende de las dimensiones K, N



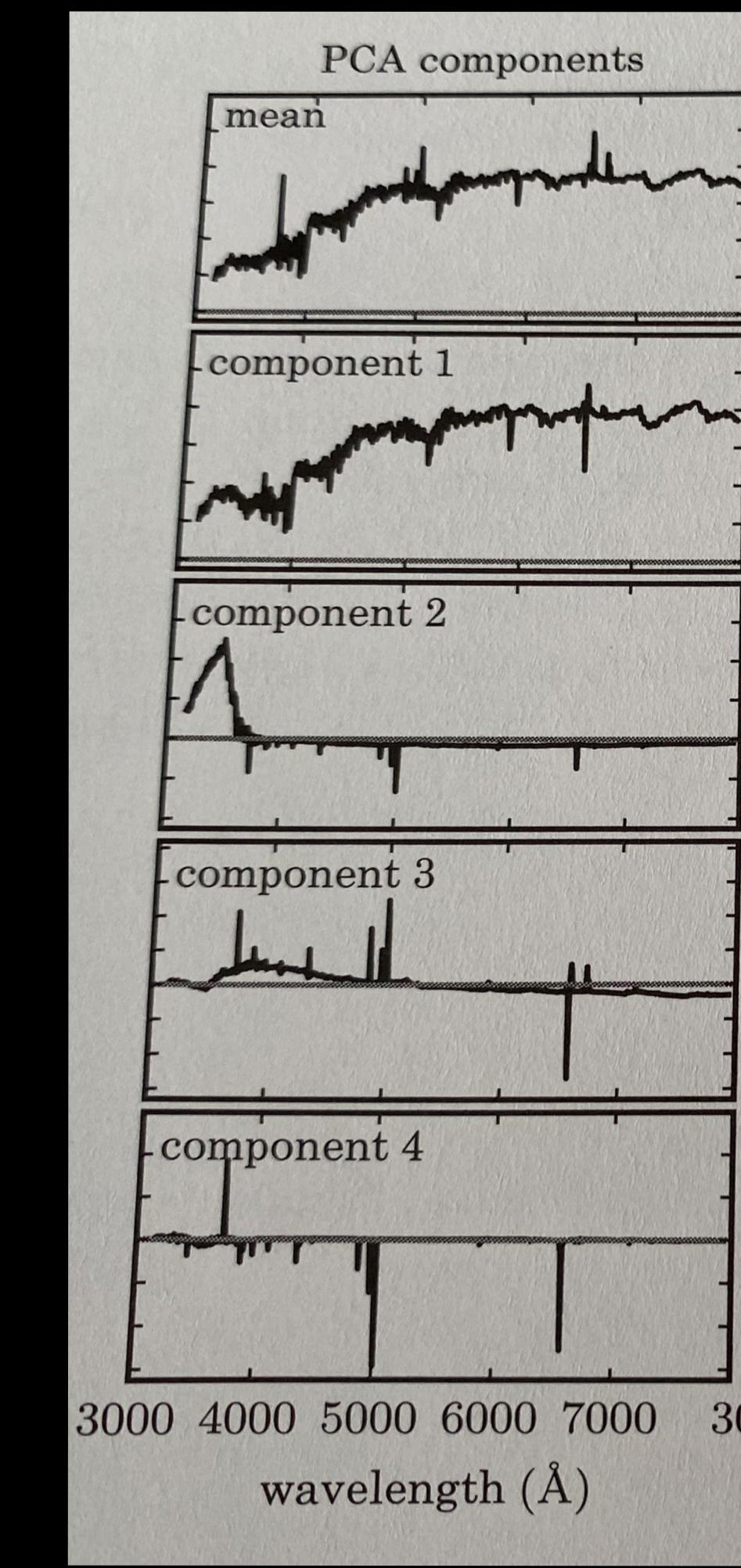
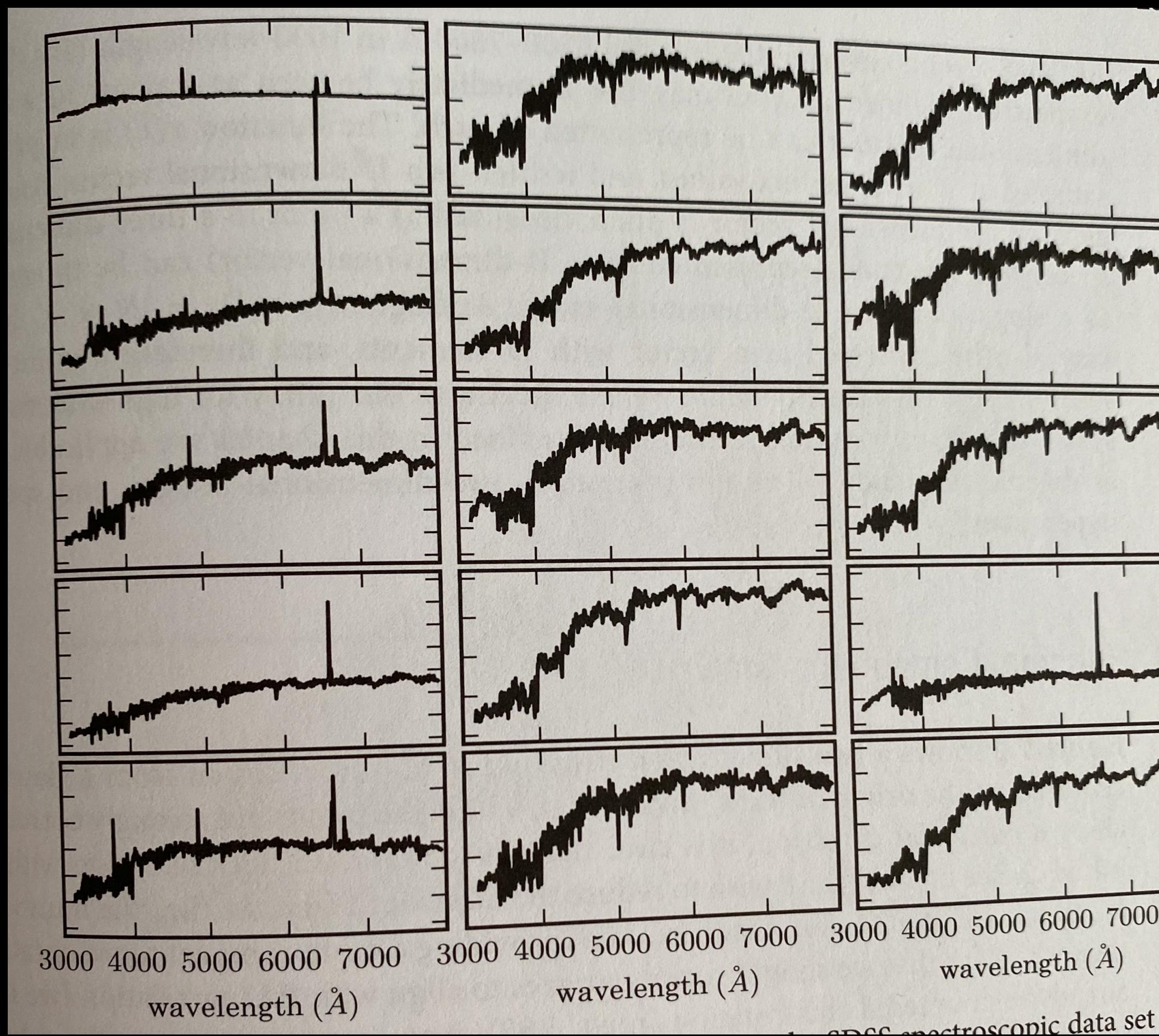
PCA

- 4) Tomar los eigenvalores y ordenarlos de mayor a menor, a la vez que construimos la matriz de eigenvectores ordenada (P^*), donde cada columna corresponde con el eigenvalor ordenado. Podemos mantener solo los primeros n-eigenvectores si ya estamos haciendo la reducción de variables.

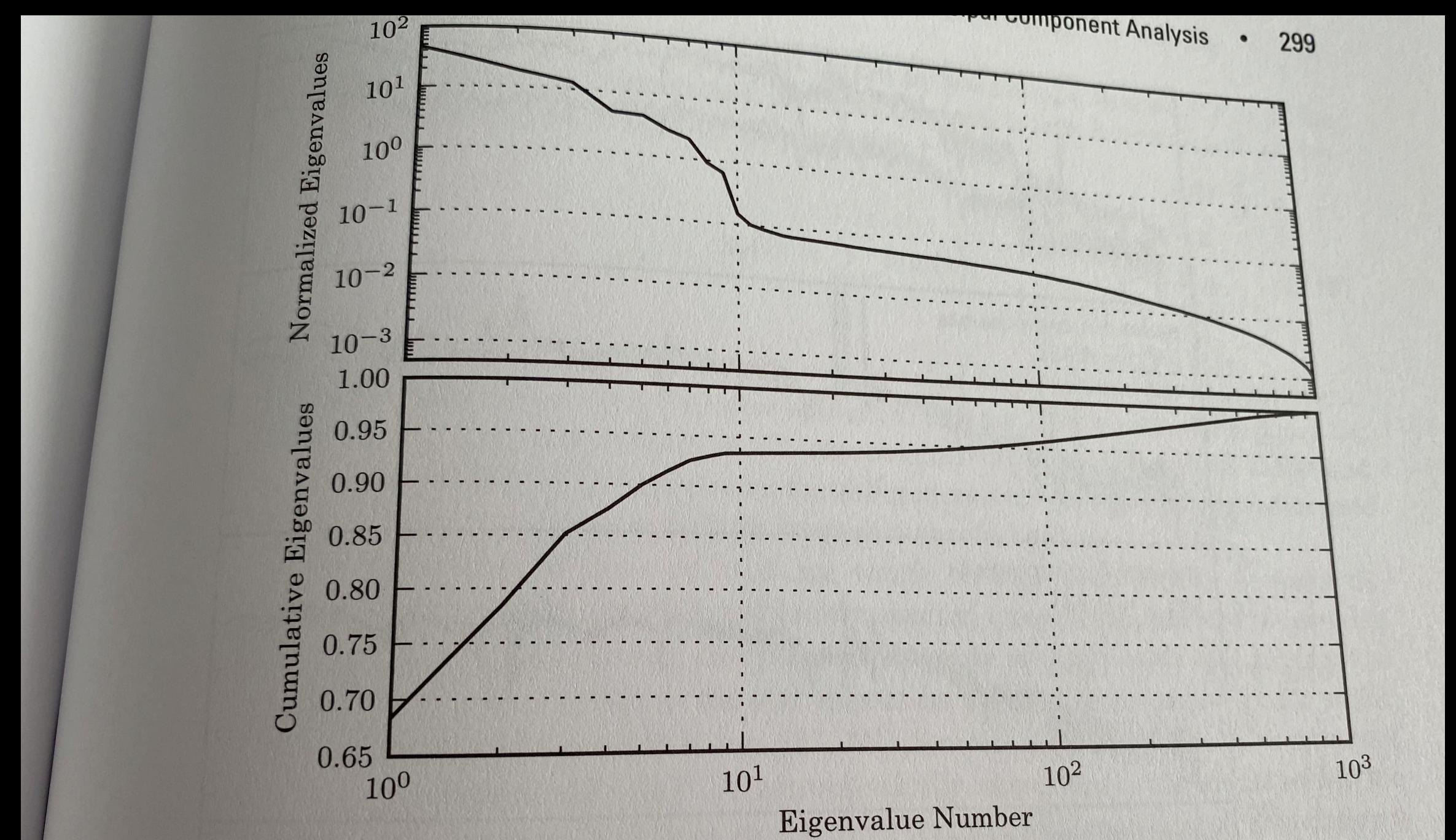
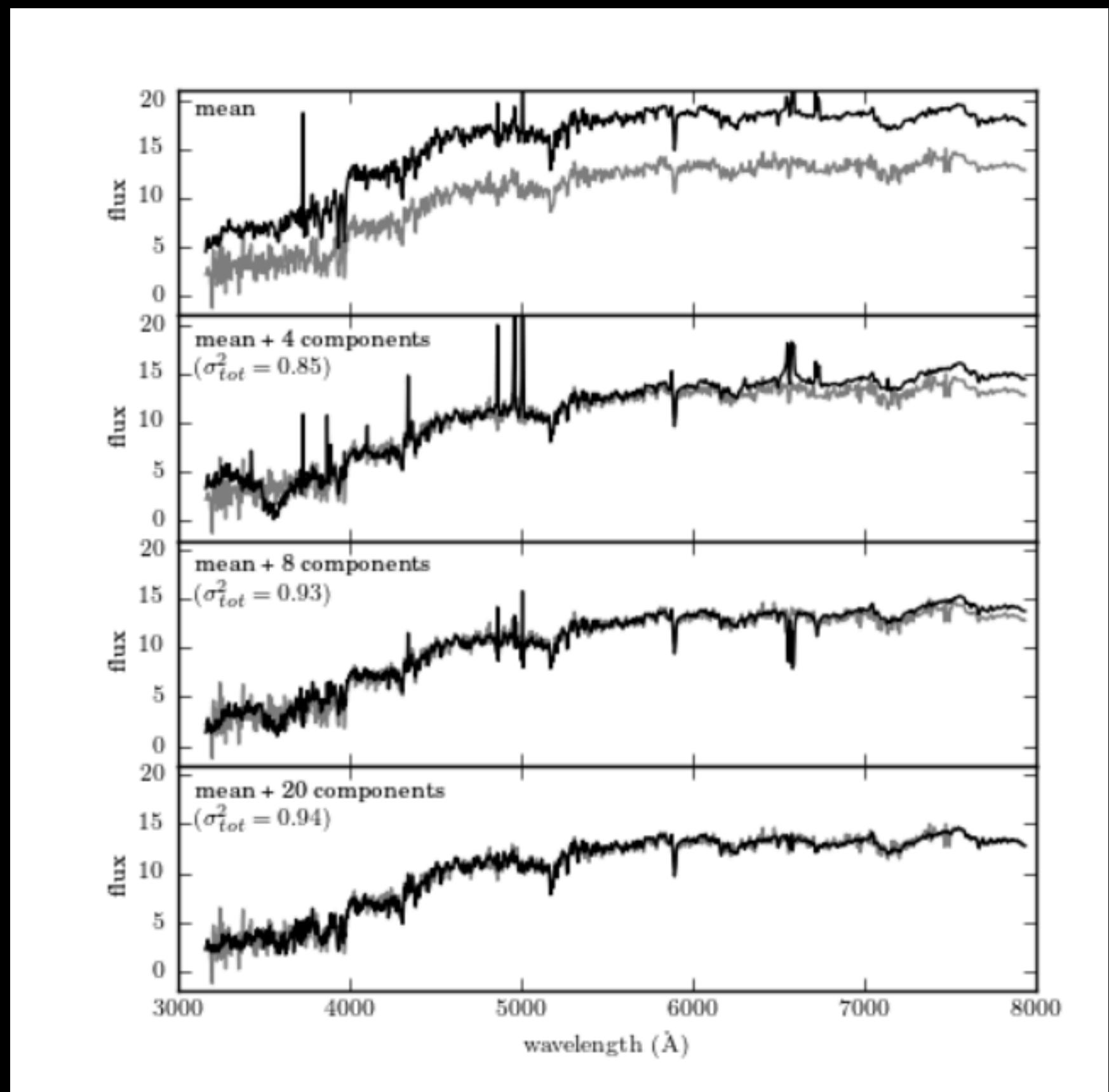
¿Qué numero de componentes es el óptimo a mantener?

- Opción 1: Determinar el número de variables que queremos manejar y descartar el resto.
- Opción 2: Determinar la proporción de varianza que se reproduce con las nuevas variables y establecer un límite .

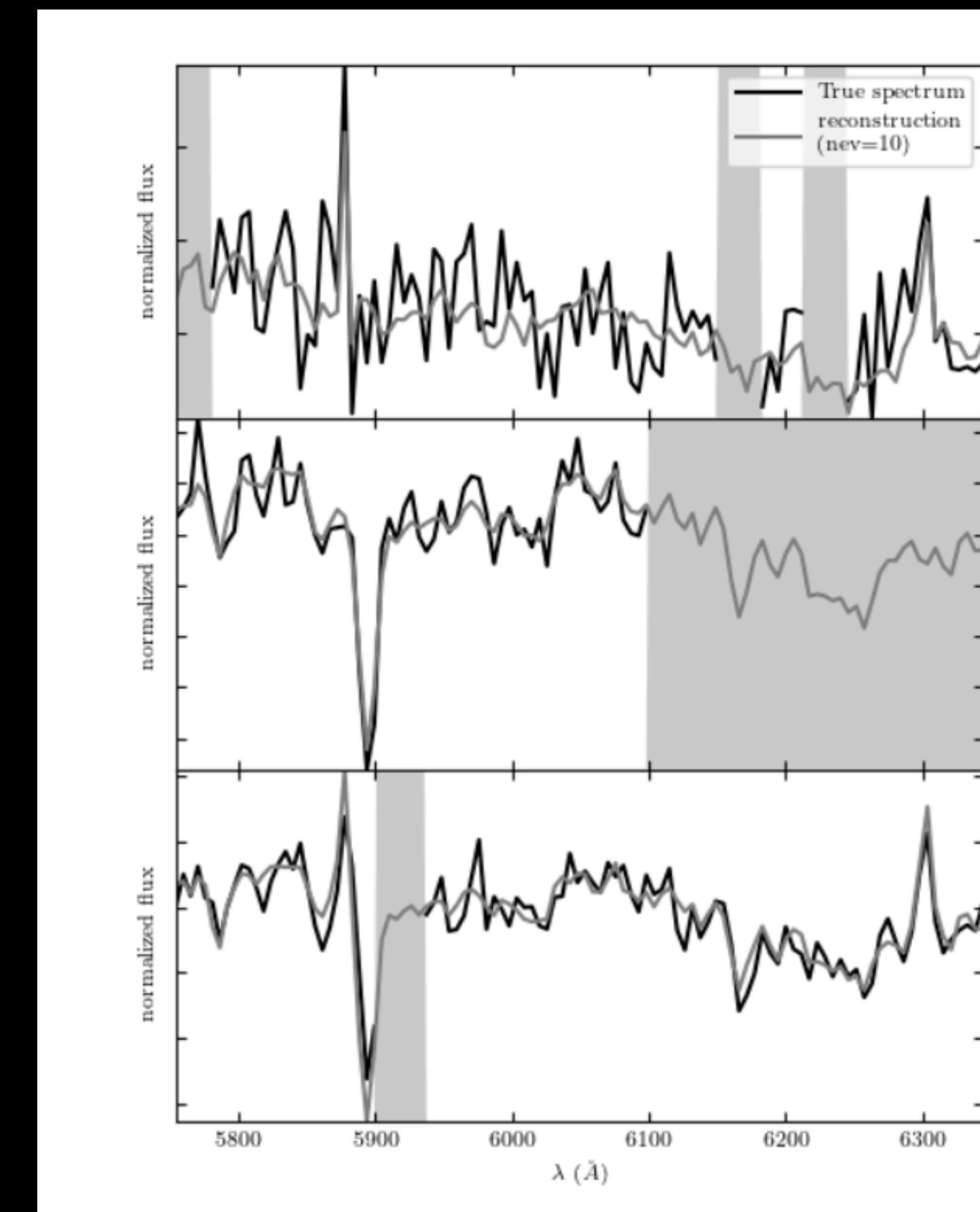
Ejemplo



Ejemplo



Ejemplo



Ejercicio

- Reproducir el ejercicio en https://www.astroml.org/book_figures/chapter7/fig_spec_reconstruction.html, i.e. la figura, y el panel izquierdo de la grafica en https://www.astroml.org/book_figures/chapter7/fig_spec_decompositions.html#book-f>ig-chapter7-fig-spec-decompositions, utilizando los datos de SDSS que vimos como accesar en el ejercicio del 2 de Octubre de 2020: eboss_qso_DR14.ipynb