Instituto Tecnológico Autónomo de México

Selección de β en Factorización de Matrices No Negativas usando la β-divergencia

Tesis
que para obtener el título de
Licenciado en Matemáticas Aplicadas
presenta

José Manuel Rodríguez Sotelo

Asesor: Dr. Manuel Mendoza Ramírez

México, D.F. 2014

Autorización

Con fundamento en el artículo 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada "Selección de β en Factorización de Matrices No Negativas usando la β-divergencia", otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Bailléres Jr. autorización para que fijen la obra en cualquier medio, incluido el elctrónico y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por la divulgación una contraprestación.

José Manuel Rodríguez Sotelo

Fecha

Firma

Para Aida

Agradecimientos

Hay estudios que sugieren que una de las claves para ser feliz es sentirse agradecido con las personas que te rodean. De confirmarse esta teoría creo que se explicaría en gran parte por que soy tan feliz. Simplemente tengo muchas razones para darle gracias a ciertas personas que han cruzado mi camino.

En primer lugar quiero darle gracias a mi familia por siempre estar ahí cuando los necesito y por consentirme. Gracias por hacerme sentir especial e inteligente. Tengo la autoestima que tengo gracias a ustedes.

Gracias Daniel por ser el mejor hermano que podría pedir. Si acaso a veces no lo parece, no te equivoques, espero cosas grandes de ti y estoy muy orgulloso de ti.

Gracias a ti, mamá Sol, por darme la formación que tengo. Gracias por ser valiente en las épocas que requerían decisiones difíciles. Gracias por trabajar cada día para darnos lo mejor a mi hermano y a mí. Gracias por siempre ponernos adelante de cualquier otra cosa. ¡Gracias! Porque no sería nadie sin ti.

Gracias a Karen y a Anelvi por resistir las dificultades de mi amistad. Se bien que no he estado siempre presente y les agradezco por ser pacientes conmigo. Espero que seamos amigos por el resto de tiempo que nos queda vivos. Espero también que no haga falta tanta paciencia de su parte; no estoy tan confiado de esto último.

Gracias a Daniel Sámano por ofrecerme mi primer trabajo formal en el Banco de México. Gracias a Claudia Ramírez por hacerme sentir en confianza. Gracias a mis amigos del Banco, María, Mau, Bobby, Miguel, Jorge, Miriam, Juan, Daniel, Diego, Luis por hacerme disfrutar mi tiempo ahí.

Gracias a Onur Dikmen por recibirme a trabajar en Finlandia y por ser guía de gran parte de este trabajo. Me siento en gran deuda con usted.

Gracias a mis maestros del ITAM por la formación envidiable que recibí. En especial muchas gracias a mis maestros de matemáticas. Al profesor César Luis por

una excelente clase de álgebra lineal. Al profesor Javier Alfaro por estar siempre al pendiente de las actividades de los estudiantes de matemáticas aplicadas. Al profesor José Luis Morales por las 3 clases más complicadas de la carrera. Estoy convencido de que soy un mejor matemático aplicado gracias a usted. A los profesores Ernesto Barrios y Alberto Tubilla por mi confiable formación en probabilidad. A la profesora Begoña por una excelente clase de estructuras. Gracias por hacerme disfrutar la programación. Particularmente, muchas gracias al profesor Preciado por una coordinación que cualquier carrera querría tener. Gracias por buscar siempre lo mejor para los estudiantes.

Muchas gracias a los profesores Luis Felipe González, Fernando Esponda e Ignacio Lobato por ser mis sinodales. No solamente sus comentarios hicieron este trabajo mejor, sino que es un honor ser evaluado por personas tan capaces como ustedes. En particular, gracias al profesor Luis Felipe por impartir la clase que definió hacia donde quiero ir ahora.

En especial me gustaría agradecer al profesor Manuel Mendoza. Además de hacer este trabajo mucho mejor con sus comentarios y su conocimiento, su ejemplo me motiva a ser un mejor científico. Espero lograr ser capaz algún día de poder inspirar a alguien como usted me inspira a mí. No estaría interesado en la estadística de no ser por usted. Es un honor para mí haber sido su estudiante.

Gracias a todos mis amigos que hicieron de mi paso en el ITAM una experiencia grata, el ITAM no hubiera sido lo mismo sin ustedes. Gracias Rafa por ser un gran ejemplo y por invitarme a Laberintos. Gracias a mis amigos matemáticos, Omar, Manuel, Mariana, Camila y Andrea. Tengo muy buenos recuerdos de tiempo con ustedes, como las cenas de Navidad. Gracias Andrea, Cris y Tania por todas las tardes de sushi y por ir al viaje.

Gracias Quique y Oscarin por ser mis amigos desde la primera semana de clases y por mantenerse así. A 6 años de nuestra primera clase de Ideas I estoy seguro de que este tiempo es solamente el inicio de una amistad duradera.

Gracias a mis amigos economistas: Imanol, Natalia, Elena, Alonso y Regina, por ñoñear conmigo. Gracias por Ixtapa, por el blog, por el GSE, y por todas las pláticas de economía y de la vida. Lo que más me gusta de la economía son ustedes.

Muchas gracias Ame, Linda, Andrea, Chunky, Raúl y Vila por todos los buenos momentos; desde jugar pictionary hasta Puerto. Cada clase era mejor con ustedes. Gracias Nick por ser mi consciencia moral cuando lo necesité. Si nos sale todo bien, nos queda mucho tiempo de trabajo juntos.

Gracias Lalo por nuestras discusiones filosóficas de viernes en las noche. Gracias por las preguntas y por las respuestas.

Gracias Andrés por ser un ejemplo a seguir. Te admiro y me encantaría ser tan valiente y aventurero como tú.

Gracias Magda por invitarme a los viajes que haces; espero tener la oportunidad de acompañarte en muchos viajes más.

Gracias Ana Cris por un excelente año de trabajo juntos; gracias por ir al viaje con todo y muletas. Muchas gracias por todas las fiestas a las que me invitaste y por ayudarme a ser más social.

Gracias Colmi por compartir conmigo tu pasión por las matemáticas y por las otras cosas que te importan. Gracias por cada partida de kuhhandel y por cada noche de alitas.

Gracias David por todos los juegos de mesa jugados, por ayudarme el día de tu examen profesional y por ser alguien con quien siempre puedo hablar de temas importantes; más de una vez me has ayudado a encontrar paz cuando la tenía perdida.

Gracias a ti, Lore, por compartir lo increíble que eres conmigo. Gracias por andar conmigo en bici. Gracias por motivarme a ser mejor. Gracias por llenar los vacíos que tenía dentro de mí. Te quiero.

Finalmente, gracias a ti, Tía Aida. Gracias por enseñarme una de las lecciones más valiosas que he aprendido; el mejor tipo de amor es el que no tiene ninguna obligación de por medio. Estaría satisfecho si con mi vida puedo servir a los demás la mitad de lo que tú has servido a los que te rodean. Eres la razón de que yo este donde estoy. Gracias.

Resumen

En este trabajo se presenta un nuevo algoritmo para seleccionar la función de costo para factorizar de forma óptima una matriz no negativa. La factorización de matrices no negativas fue popularizada por Lee y Seung [26] en la comunidad de aprendizaje de máquina por sus propiedades para aprender estructuras basadas en partes; esta característica la volvió atractiva para ser usada en diferentes aplicaciones como:

- Clasificación de noticias en temas.
- Identificar rasgos faciales comunes.
- Extracción de patrones en canciones.
- Agrupación de canciones por el tema de sus letras.

Adicionalmente, se presenta el algoritmo de aprendizaje no supervisado más conocido y usado en la actualidad, el análisis de componentes principales. Este método se presenta con la finalidad de tener una base para explicar y resaltar las características de la factorización de matrices no negativas. Estas dos herramientas de análisis se utilizan para resulver dos problemas muy parecidos.

La elección de la función de costo se restringe a la familia de las β-Divergencias. Para la selección de la función dentro de esta familia se plantea un modelo estadístico basado en la familia de distribuciones Tweedie, que forma parte de los modelos de dispersión exponencial. Usando dicho modelo estadístico, se plantea la posibilidad de elegir la función de pérdidas maximizando la verosimilitud del parámetro correspondiente.

Índice general

Índice de figuras

Capítulo 1

Introducción

> "Enseñar no es una función vital
> porque no tiene el fin en sí misma;
> la función vital es aprender."
> Aristóteles

Llegar a una conclusión sobre lo que significa "aprender" sin lugar a dudas entusiasmaría a tu filósofo de confianza. El concepto de aprendizaje comprende procesos tan diferentes que es difícil de definir con precisión. Algunas posibles definiciones de *aprender* son: "Adquirir conocimiento de algo por medio de estudio o de experiencia", "Incorporar algo en la memoria","Modificar la tendencia de comportamiento basado en experiencia".

En este trabajo se presenta un nuevo algoritmo para seleccionar la función de costo en el método de factorización de matrices no negativas. La factorización de matrices no negativas (FMN) fue popularizada por Lee y Seung [26] en la comunidad de aprendizaje de máquina por sus propiedades para aprender estructuras basadas en partes. La FMN forma parte del conjunto de técnicas conocido como *Aprendizaje de máquina*.

1.1. Aprendizaje de Máquina

A lo largo del tiempo, se ha buscado conseguir que las máquinas aprendan. Como Russell y Norvig [35] describen, durante los primeros años de la inteligencia artificial (1952 - 1969) se llegó a pensar que en pocos años las máquinas serían capaces de superar a los seres humanos en algunas tareas particularmente complejas. Sin em-

bargo, este optimismo llego a su fin debido a las limitaciones computacionales de ese tiempo. Incluso se abandonó la esperanza de que una computadora sería capaz de traducir un idioma en tiempo real, o que se crearía una inteligencia artificial capaz de mantener una conversación con una persona.

A pesar del atraso por las limitaciones en la capacidad de cómputo de la época, esos años fueron muy fructíferos en cuanto a la cantidad de ideas sobre el aprendizaje que se generaron. En particular, se adaptaron teorías sobre el aprendizaje generadas en otros campos del conocimiento como la psicología. El origen de algunas de las técnicas diseñadas para que las máquinas puedan aprender, está en los esfuerzos de psicólogos para crear modelos que puedan representar el aprendizaje humano y animal.

Por otra parte, desde hace algunos años, la cantidad de datos disponibles para efectos de análisis ha crecido de forma impresionante. La comunidad científica ha respondido con algoritmos que permiten extraer información valiosa de los datos. El campo de procesamiento de señales fue de los primeros en encontrarse al problema de extraer información de una base de datos, y de este campo surgieron algunos de los algoritmos más populares.

Debido a la creciente importancia de los problemas relacionados con la extracción de información, el estudio de los diferentes algoritmos se ha consolidado en un campo emergente conocido como "Aprendizaje de Máquina". De acuerdo con Nilsson [30], una máquina aprende siempre que cambia su estructura, programas, o datos (basada en insumos o en respuesta a información externa) de tal manera que su desempeño futuro tiende a mejorar. Nilsson comenta que el aprendizaje de máquina normalmente se refiere a cambios en sistemas que realizan actividades asociadas al campo de Inteligencia Artificial. Esas labores involucran reconocer patrones, elaborar diagnósticos, controlar robots, predecir variables, entre otras.

Los avances y las ideas en el aprendizaje de máquina han provenido de diferentes disciplinas. Cada una con sus propios problemas, métodos y notaciones. Algunas de las que han contribuido son:

- Estadística
- Computación

- Inteligencia Artificial
- Investigación de Operaciones
- Optimización Numérica
- Neurociencia
- Psicología

El campo del aprendizaje de máquina se ocupa de diversos temas y es muy heterogéneno en lo que respecta a las perspectivas con que se ha aproximado a las soluciones de los problemas. Una complicación que vale la pena tener en cuenta es que la notación suele ser diferente según la perspectiva desde la cual se plantee un problema.

1.2. Aprendizaje Estadístico

La estadística es constantemente desafiada por diferentes ramas de la ciencia, así como por la industria. Tiempo atrás, los problemas solían provenir de experimentos relativamente pequeños de la agricultura y de la industria. Dado el desarrollo de los sistemas de información y del poder de cómputo disponible, los problemas se han hecho cada vez más grandes y complejos. Hoy en día se generan volúmenes de datos impresionantes. Usando las técnicas estadísticas existendes se puede *aprender de los datos*; esto es, extraer patrones y tendencias, y entender lo que los datos "dicen". El *Aprendizaje Estadístico* se refiere a un amplio conjunto de herramientas que sirven para entender los datos.

Tradicionalmente, una de las labores que los estadísticos han enfrentado es resolver problemas que requieren extraer resúmenes de información contenida en datos. Para lograr hacer inferencia de los datos se suelen plantear modelos estadísticos que representen el estado subyacente del mundo que generó ciertos datos y después extraer parámetros clave. La característica que hace diferente al aprendizaje estadístico del aprendizaje de máquina es la búsqueda de un modelo de probabilidad subyacente dentro de una conjunto de datos. Con este modelo se busca lograr representar la variabilidad que existe en los datos.

A grandes rasgos, todas las técnicas de aprendizaje estadístico involucran la construcción de un modelo estadístico para predecir un resultado, basado en uno o más

insumos o para descubrir patrones en los datos. Los problemas que se pueden abordar usando este enfoque se caracterizan por describir fenómenos que se manifiestan a través de datos que presentan variabilidad e incluyen diversos campos como física, economía, medicina, finanzas y muchos más.

1.2.1. Historia del Aprendizaje Estadístico

Aunque el término "aprendizaje estadístico" es relativamente nuevo, muchas de las técnicas que forman parte de este campo se desarrollaron mucho tiempo atrás. Al inicio del siglo XIX, se desarrolló el método de mínimos cuadrados que es el primer método de ajuste de curvas y que se encuentra en el origen de la regresión lineal. Para el problema de clasificación en grupos, Fischer propuso el análisis de discriminante lineal más de un siglo después en 1936 y en 1940 se propuso la regresión logistica. Para 1980 el poder de cómputo había mejorado lo suficiente que se pudieron explorar métodos no lineales. Así, en 1984 se tenía por primera vez una implementación de los árboles de clasificación y regresión de la mano de Breiman et al. [3]. En los últimos años se han generado distintos métodos que complementan y generalizan los métodos existentes. Actualmente, las técnicas de aprendizaje estadístico se encuentran en expansión con nuevas ideas que prometen mejorar los límites actuales en la predicción y descubrimiento de patrones en bases de datos cada día más desafiantes. La mejor introducción a este campo se encuentra en James et al. [21] que complementa el trabajo anterior de Hastie et al. [18] que se dirige a un público más técnico.

1.3. Aprendizaje Supervisado

El problema típico de aprendizaje supervisado consiste en predecir el valor de una variable respuesta Y dado un conjunto de variables de entrada $X^T = (X_1, \ldots, X_p)$. Las predicciones estan basadas en un conjunto de entrenamiento y se suele referir a este tipo de aprendizaje como "Aprender con un maestro". La metáfora detras de esta expresión es que el modelo o "estudiante" presenta una respuesta $\hat{y}$ para cada $x = x_1, \ldots, x_p)$ del conjunto de entrenamiento, y el "maestro" provee la respuesta corecta y/o un error asociado a la respuesta del estudiante. Esto suele ser caracterizado por una función de pérdida L que penaliza los errores cometidos por el "estudiante".

1.4. Aprendizaje No supervisado

El aprendizaje no supervisado consiste de un conjunto de herramientas enfocadas a la situación en la que solamente se cuenta con un conjunto de n observaciones de las variables $X_1, X_2, \ldots, X_p$. En este contexto, no se puede hablar de predicción dado que no hay variables respuestas que predecir. En lugar de eso, el objetivo es descubrir patrones interesantes o estructura en los datos con los que se cuenta. En general, es un campo menos desarrollado que el aprendizaje supervisado, aunque no por eso menos importante.

El aprendizaje supervisado es un campo con un objetivo claro, predecir de la mejor forma variables respuesta a partir de observaciones de una colección de variables. Más aun, existe una forma clara de cuantificar la calidad de los resultados obtenidos: Evaluando la predicción en un conjunto de prueba que no fue usado para entrenar el modelo.

El aprendizaje no supervisado es mucho más desafiante. En este campo existe mucha más subjetividad y no hay una meta clara para los análisis. Como consecuencia, puede llegar a ser muy difícil evaluar la calidad de los resultados obtenidos. Dicho de otra forma, no hay forma de saber si el trabajo que se realiza está bien debido a que la respuesta correcta es desconocida, si es que esta existe.

Algunas de las preguntas que se pueden explorar usando técnicas de aprendizaje no supervisado son: ¿Qué tipo de visualizaciones podemos usar para extraer información de los datos? ¿Existen subgrupos dentro de las variables o dentro de las observaciones de los datos? ¿Se pueden encontrar observaciones dentro de la base que puedan caracterizar subgrupos en los datos?.

También vale la pena destacar que es común dentro de un análisis estadístico usar las técnicas de aprendizaje no supervisado disponibles como parte del proceso de exploración de datos después hacer uso de los resultados de los análisis realizados para intentar predecir más eficazmente alguna variable de interés; un ejemplo muy común de esto es cuando se usan los componentes principales en una regresión.

1.5. Bases de Datos

A continuación se presenta una descripción breve de las bases de datos utilizadas en este trabajo. Con estas bases de datos se busca que la explicación de los algoritmos

presentados sea clara y suficiente para reconocer las diferencias entre las diferentes alternativas.

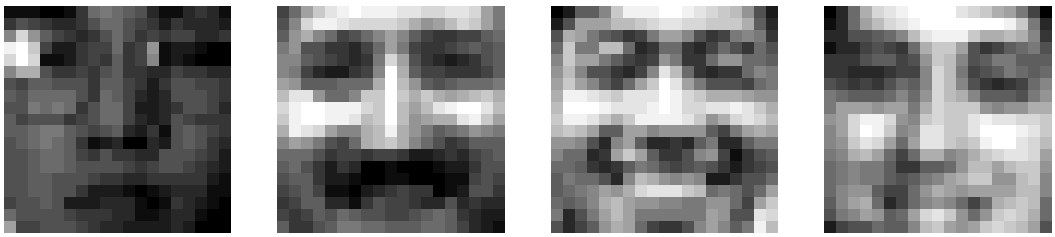


Figura 1.1: Ejemplos de imágenes de caras.

1.5.1. Caras

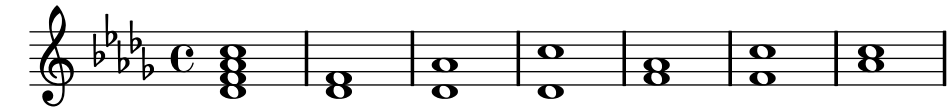
Esta base consiste de 2,429 imágenes de caras de baja resolución usadas por Lee y Seung [26]. La resolución de cada imagen es de 19×19 pixeles en escala de grises. En la figura 1.1 se muestran algunos ejemplos del tipo de imágenes de las que consiste la base de datos.

1.5.2. Letras de Canciones

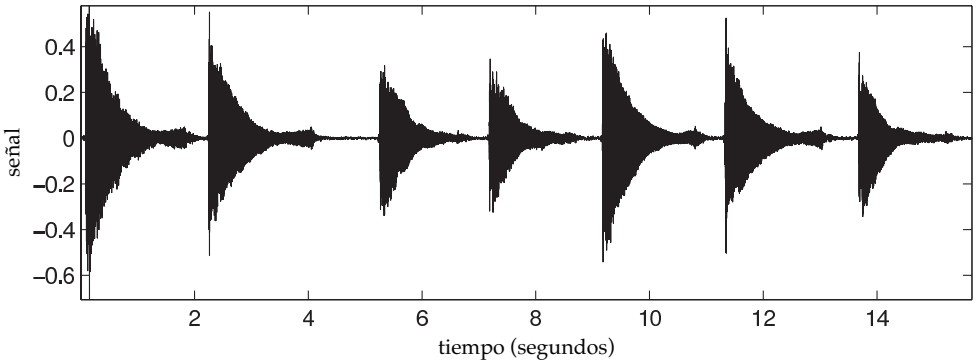
Esta base consiste de conteos de palabras en la letra de 237,701 canciones. En total, existen 498,134 palabras únicas y 55,163,335 ocurrencias de estas palabras. Por simplicidad, sólo se analizaron las 5,000 palabras que ocurren más frecuentemente en las canciones. Estas 5000 palabras aparecen 50,607,582 veces, con lo cual representan aproximadamente el 92 % de las palabras en total. Esta base fue construida por Bertin-Mahieux et al. [2] y es reconocida como la base de letras de canciones más grande y limpia disponible para investigación.

1.5.3. Piano

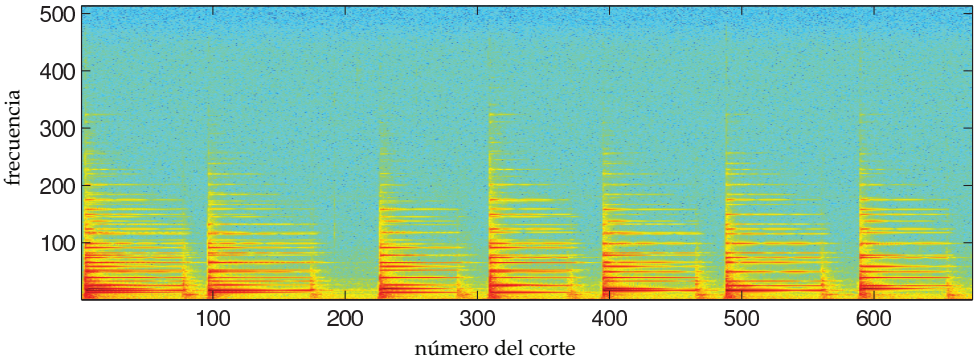
Los datos provienen de la grabación de un piano Yamaha DisKlavier MX100A ubicado al tocar 4 notas; primero todas al mismo tiempo, y después por pares en todas las combinaciones posibles. La grabación duró 15.6 segundos y se construye el espectrograma de la grabación usando la transformada de Fourier de tiempo reducido, que consiste de una matriz que contiene la intensidad de distintas señales en el tiempo. Se consideran 513 diferentes frecuencias en 674 cortes en el tiempo. En



(a) Notas musicales



(b) Señal



(c) Espectrograma

Figura 1.2: Distintas representaciones de los datos de sonido.

la figura 1.2 se presentan las notas, la señal en el tiempo y su espectrograma. Para más información sobre estas y otras transformaciones comunes asociadas a datos provenientes de audio se puede consultar Wang [36].

Capítulo 2

Análisis de Componentes Principales

En este trabajo se presenta el algoritmo conocido como factorización de matrices no negativas. Con el propósito de explicar a detalle el funcionamiento de este algoritmo, se le compara con el algoritmo más conocido de aprendizaje no supervisado: El análisis de componentes principales (ACP). El objetivo de este capítulo es detallar el análisis de componentes principales con el propósito de comparar estos dos algoritmos de aprendizaje; ilustrando sus similitudes y enfatizando sus diferencias. La comparación se realizará en el contexto de la base de imágenes de caras que se presentó en la sección 1.5.1.

El análisis de componentes principales es un método estadístico que sirve para producir una combinación lineal y reducir la dimensionalidad de los datos. El análisis de componentes principales presume que hay variables relacionadas entre sí en la base de datos. La idea es reducir la dimensión en la representación de los datos intentando mantener toda la variación posible. Esto se logra transformando las variables de la base de datos, a un nuevo conjunto de variables llamadas 'Componentes Principales' (CPs). Las restricciones sobre las componentes son que sean no correlacionadas entre sí y ordenadas de tal forma que con pocas de estas variables se pueda retener la mayoría de la variación contenida en todas las variables iniciales.

2.1. Combinación Lineal

En la base de datos, cada imagen tiene 361 pixeles. Así, se puede pensar que cada imagen representa un vector en un espacio de dimensión 361. Una pregunta natural que surge con esta interpretación es ¿Cuál es la distribución de estas imágenes en dicho espacio? Claramente, la distribución no es uniforme puesto que si simulamos

una distribución uniforme en dicho espacio y observamos los resultados como imagen, solamente tendríamos una imagen que luce como "ruido". Por lo tanto, existe cierta estructura en la distribución de estas imágenes sobre el espacio de dimensión 361 que las hace cualitativamente diferentes al "ruido".

¿Cómo podemos modelar esta estructura? Una alternativa es imaginar que cada imagen está formada por una combinación lineal de "características" (comunes para todas las imágenes de la base). Cada "característica" es a su vez un vector de dimensión 361 y puede ser interpretado como una imagen del mismo tamaño que las imágenes. Si v es una imagen y $W_1, ..., W_K$ son "características", entonces:

$$v \approx \sum_{k=1}^{K} H^k W_k$$

donde H^k son los pesos con los que las "características" se combinan. Dicho de otra forma, $H = (H^1, ..., H^K)$ son los pesos de la combinación lineal de v en la base de "características". Los pesos H^k con los que se combinan las "características" son diferentes para cada imagen mientras que las "características" $W_1, ..., W_K$ son comunes para todas las imágenes de la base.

2.2. Planteamiento del Problema

Una de las razones de la popularidad de este algoritmo es que a lo largo del tiempo se ha conseguido plantear desde diferentes perspectivas. En este capítulo, se mostrará la versión de componentes principales como un problema de factorización de matrices planteada por Nikolov [29], así como la versión algebraica que se presenta en Jolliffe [22].

2.2.1. Factorización de Matrices

Dada una matriz V de dimensión $F \times N$ el problema de descomposición en componentes principales consiste en encontrar una factorización:

$$V \approx WH = \hat{V}$$

donde W y H son matrices de dimensión $F \times K$ y $K \times N$ respectivamente. La idea es que K sea pequeña de tal forma que $\hat{V}$ sea una matriz de bajo rango, con pocos parámetros. En este contexto, pequeña significa que $K \times (F + N) \ll FN$. En lo siguiente v_{fn}, w_{fk}, h_{kn} y $\hat{v}_{fn}$ denotan las entradas de las matrices en cuestión.

El problema de optimización asociado al problema de factorización es el siguiente:

$$\begin{aligned} \underset{W,H}{\text{minimizar}} \quad & D(V||WH) = \frac{1}{2}\sum_{f=1}^{F}\sum_{n=1}^{N}(v_{fn} - \hat{v}_{fn})^2 \\ \text{sujeto a} \quad & W^TW = I \end{aligned}$$

En donde la restricción $W^TW = I$ establece ortogonalidad entre las diferentes componentes principales de los datos. Es importante señalar que la cantidad de características a considerar K se fija de antemano. Este problema es importante por sí mismo, en este trabajo se asume que se tiene una forma de seleccionar este parámetro.

Vale la pena destacar que cada columna de W es una de las "características" que se describieron anteriormente. Por otro lado, cada renglón de H contiene los pesos de la combinación lineal con los cuales se deben ponderar las "características" para generar una aproximación a la imagen inicial correspondiente.

2.3. Historia

Aunque existen trabajos previos que detallan la descomposición de una matriz en valores singulares de tal forma que se podría derivar el ACP, es generalmente aceptado que los primeros trabajos que describen el ACP son los de Pearson [32] y Hotelling [19]. Estas dos publicaciones adoptan diferentes enfoques para el planteamiento de este problema. La derivación de Hotelling es muy parecida a la versión algebraica moderna que se presenta en Jolliffe [22]. Pearson, por otro lado, con un enfoque geométrico estaba interesado en encontrar líneas y planos que ajustaran de la mejor forma posible a un conjunto de puntos en un espacio multidimensional.

2.4. Resultados

El análisis de componentes principales se ha utilizado exitosamente en diferentes aplicaciones. Entre estas se incluyen, creación de números índice, como por ejemplo índices de pobreza y compresión de información. También se emplea en la selección de variables en los modelos demograf regresión. Los campos en los que se ha aplicado en ACP incluyen agricultura, biología, demografía, ecología, economía, genética, geología y química, por ejemplo.

2.4.1. Caras

El ACP se utiliza en particular para la compresión de imágenes. En la figura 2.1 se observa la reconstrucción de imágenes que se logra usando las primeras 40 componentes principales, esto significa una reducción del tamaño de la base de 90 %. Aun después de reducir la dimensión de los datos, cada una de las caras en el ejemplo se puede distinguir con razonable facilidad.

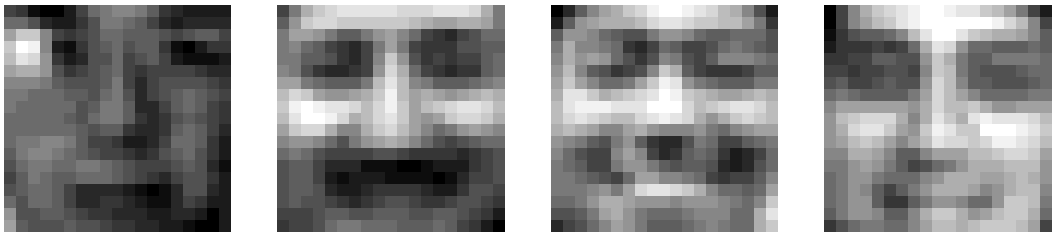


Figura 2.1: Reconstrucción de las imágenes usando ACP con $k = 40$

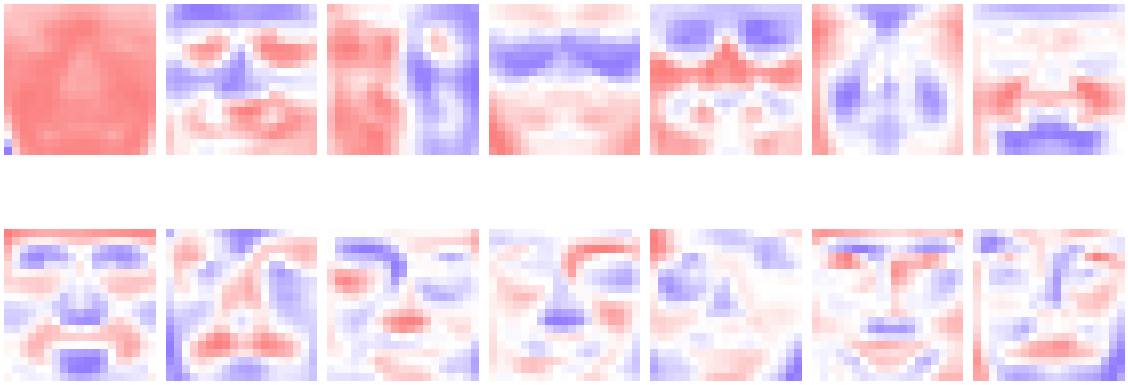


Figura 2.2: Primeras componentes principales cuando $k = 40$.

La representación que se obtiene con un análisis de componentes principales tiene la característica de que cada componente usa todas las variables de la base, es decir, una representación hólistica. Esta característica tiene puntos a favor y en contra. Al usar todas las variables, se puede lograr una representación eficiente en pocas dimensiones. Por otro lado, esto dificulta la interpretación de los resultados. La figura

2.2 muestra una representación de las primeras 12 componentes principales de las 40 que se utilizan para reconstruir las imágenes en el ejemplo. Lo más importante, es que cada componente principal usa todos los pixeles y que tienen variables con signos diferentes (los signos se representan con los colores de la imagen). La recomposición de cada imagen se consigue con una combinación lineal de las primeras 40 componentes.

Capítulo 3

Factorización de Matrices No Negativas

"El todo es mayor a la suma de las partes"

Aristóteles

La factorización de matrices no negativas fue popularizada por Lee y Seung [26] con su artículo en la Revista Nature en 1999 al resaltar que los resultados de este algoritmo permiten un 'aprendizaje basado en partes'. Esto es completamente diferente al 'aprendizaje holístico' de otros algoritmos como el ACP. Este resultado es importante, debido a la intuición filosófica de que un objeto o concepto se puede representar como la suma de sus partes. La suma es una, entre muchas, formas de combinar las partes.

3.1. Combinacion Lineal Aditiva

Igual que en el caso del ACP, la pregunta que se intenta solucionar en la FMN es ¿Cómo modelar la estructura subyacente de las imágenes de forma simplificada? La factorización de matrices no negativas se caracteriza por aproximar cada imagen por una combinación lineal aditiva de "características".

Además, debido a que los pesos con los que se combinan estas "características" son siempre positivos, se puede interpretar fácilmente esta combinación lineal. Una posible interpretación de la combinación lineal equivale a pensar que para construir cada cara se pegan estampas transparentes de las características hasta construir una aproximación a la imagen inicial.

Si v es una imagen y $W_1, ..., W_K > 0$ son "características", entonces:

$$v \approx \sum_{k=1}^{K} H^k W_k$$

donde $H^k > 0$ son los pesos con los que las "características" se suman. Dicho de otra forma, $H = (H^1, ..., H^K)$ son los pesos de la combinación lineal de v en la base de "características".

3.2. Planteamiento del Problema

Dada una matriz V de dimensión $F \times N$ con entradas no negativas, el problema de factorización de matrices no negativas consiste en encontrar una factorización:

$$V \approx WH = \hat{V}$$

donde W y H son matrices con entradas no negativas de dimensión $F \times K$ y $K \times N$ respectivamente. Igual que antes, la idea es que K sea pequeña de tal forma que $\hat{V}$ sea una matriz de bajo rango. Las restricciones de no negatividad que se imponen en este problema causan que se obtenga una representación basada en partes porque solamenten se permiten operaciones aditivas no sustractivas.

3.2.1. Problema de Optimización

Este problema se plantea como uno de optimización:

$$\begin{aligned} \underset{W,H}{\text{minimizar}} \quad & D(V||WH) = D(V||\hat{V}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d\left(v_{fn} \mid \hat{v}_{fn}\right) \\ \text{sujeto a} \quad & W, H \geq 0 \end{aligned}$$

en donde $d(x|y)$ es una función escalar de discrepancia o costo que penaliza la diferencia entre x y y.

La función $d(x|y)$ juega un papel muy importante en el problema de optimización. En el capítulo 4 se presentan la familia de β-divergencias que suelen ser usadas en la práctica en la factorización de matrices no negativas. Los resultados que se obtienen de la factorización dependen de la función de costo usada por lo cual es importante usar una función que adecuada para el problema determinado. Para el resto del capítulo, se asume que $d(x|y)$ es la divergencia Euclideana, que se presenta formalmente en la sección 4.1.1.

3.3. Resultados

A continuación se presentan dos aplicaciones en donde ha sido usado este algorimo. Los métodos para solucionar numéricamente este problema se detallan en el apéndice A. Estos métodos fueron desarrollados por Lee y Seung [27] y posteriormente han sido extendidos por diversos investigadores como Févotte y Idier [16], Li et al. [28], Dhillon y Sra [8] entre otros.

3.3.1. Caras

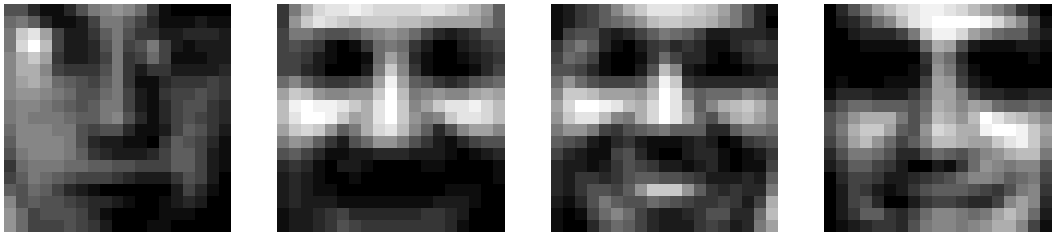


Figura 3.1: Reconstrucción de las imágenes usando FMN con $k = 40$

En la figura 3.1 es posible apreciar que la recomposición de las imágenes es de una calidad ligeramente menor que usando componentes principales. Sin embargo, la recomposición lograda tiene una calidad razonablemente parecida a la del ACP.

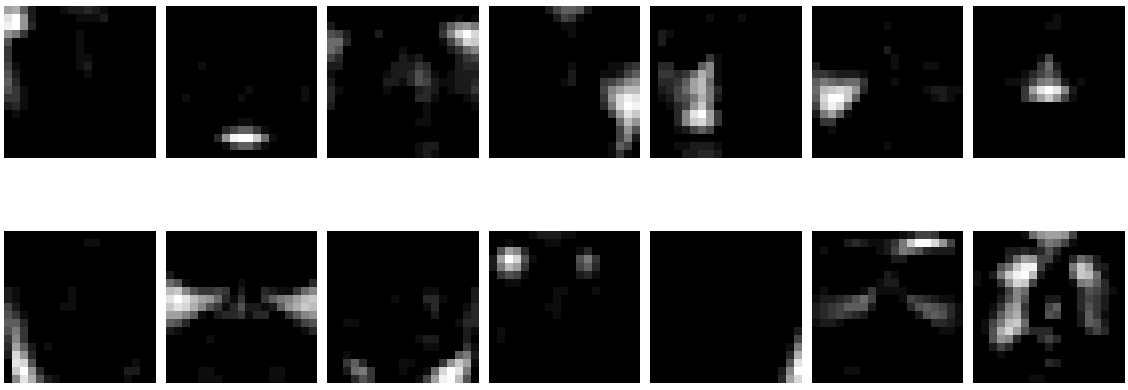


Figura 3.2: Características aprendidas por FMN con $k = 40$.

Lo más importante en la figura 3.2 es que las características identificadas usando FMN usan sólo un subconjunto de variables, a diferencia de la representación holística de componentes principales. En algunas "características" podemos observar partes de caras como bocas, ojos o narices, mientras que en la figura 2.2 cada característica parece una cara completa.

3.3.2. Letras de Canciones

A continuación se presentan los resultados de aplicar la FMN a la base de datos de letras de canciones presentada en la sección 1.5.2. En las tablas 3.1 y 3.2, cada característica (columna) está representada por las 15 palabras más importantes. El último renglón corresponde al peso con el cual se pondera cada columna. Cabe destacar que estas 8 columnas corresponden al 68 % y 83 % de las canciones respectivamente.

Cuadro 3.1: Características más importantes para la canción "Do You Love Me?" de Nick Cave and the Bad Seeds.

the	you	love	not	i	me	she	god
in	your	buy	do	am	give	her	of
and	can	liar	wanna	myself	tell	girl	blood
of	if	tender	care	like	call	beauti	soul
world	know	dear	bad	know	mmm	woman	death
with	want	instrument	nobodi	need	show	&	die
they	make	mood	anyth	want	beg	queen	fear
from	when	treasur	want	feel	rescu	sex	pain
as	see	emot	worri	caus	teas	sexi	hell
by	yourself	untru	ai	and	squeez	cloth	power
at	need	surrend	treat	out	everytim	herself	within
out	with	deeper	but	sorri	knee	doll	shall
to	feel	sparkl	know	see	strife	shes	earth
into	that	sweetest	money	in	contempl	pink	blind
sky	how	diamond	hurt	swear	guarante	gypsi	human
0.15	0.10	0.09	0.08	0.08	0.08	0.06	0.04

Cuadro 3.2: Características más importantes para la canción "California Love" de 2pac.

get	the	shake	it	you	we	come	yeah
nigga	in	motion	is	your	our	babi	five
the	and	bump	take	can	us	magic	four
ya	of	groov	doe	if	togeth	til	woo
shit	world	booti	make	know	both	lovin	summertim
like	with	shakin	easi	want	higher	im	girlfriend
fuck	they	thigh	matter	make	ourselv	shi	wow
em	from	oon	real	when	each	bodi	lala
got	as	shiver	game	see	divid	sweat	grip
hit	by	panic	possibl	yourself	nation	cant	pine
bitch	at	dick	play	need	unit	your	engin
up	out	claw	chanc	with	other	birthday	feather
off	to	opportun	give	feel	noel	wont	clap
yall	into	collid	harder	that	standard	there	mornin
ass	sky	ness	quit	how	rule	bella	gotta
0.49	0.09	0.08	0.07	0.03	0.03	0.02	0.02

Cuadro 3.3: Canciones más representativas para ciertas "características" y palabras que aparecen más frecuentemente.

($k = 2$) get nigga the ya shit like fuck em got hit bitch up off yall ass they that cmon money and	
UGK (Underground Kingz) - Murder	i the to nigga my a you got murder and it is am from we so with they yo cuz
Big Punisher - Nigga Shit	shit that nigga the i and my what to out am in on for love me with gettin you do
E-40 - Turf Drop [Clean]	gasolin the my i hey to a it on you some fuck spit of what one ride nigga sick gold
Cam'Ron - Sports Drugs & Entertainment	a the you i got yo stop shot is caus or street jump short wick either to on but in
Foxy Brown - Chyna Whyte	the nigga and you shit i not yall to a on with bitch no fuck uh it money white huh
($k = 8$) god of blood soul death die fear pain hell power within shall earth blind human bleed scream evil holi peac	
Demolition Hammer - Epidemic Of Violence	of pain death reign violenc and a kill rage vicious the to in down blue dead cold
Disgorge - Parallels Of Infinite Torture	of the tortur by their within upon flow throne infinit are no they see life eye befor
Tacere - Beyond Silence	silenc beyond a dark beauti i the you to and me it not in my is of your that do
Cannibal Corpse - Perverse Suffering	to my pain of i me for agoni in by and from way etern lust tortur crave the not be
Showbread - Sampsa Meets Kafka	to of no one die death loneli starv i the you and a me it not in my is your
($k = 26$) she her girl beauti woman & queen sex sexi cloth herself doll shes pink gypsi bodi midnight callin dress hair	
Headhunter - Sex & Drugs & RockÑ Roll	& sex drug rock roll n is good veri inde and not my are all need dead bodi brain i
Holy Barbarians - She	she of kind girl my is the a littl woman like world and gone destroy tiger me on an
X - Devil Doll	devil doll her she and a the in is of eye bone & shoe rag batter you to on no
Kittie - Paperdoll	her she you i now soul pain to is down want eat fit size and not in all dead bodi
Ottawan - D.I.S.C.O.	is she oh disco i o s d c super incred a crazi such desir sexi complic special candi
($k = 13$) je et les le pas dan pour des cest qui de tout mon moi au comm ne sur jai	
Veronique Sanson - Feminin	cest comm le car de bien se les mai a fait devant heur du et une quon quelqu etre
Nevrotic Explosion - Heritage	quon faut mieux pour nous qui nos ceux de la un plus tous honor parent ami oui
Kells - Sans teint	de la se le san des est loin peur reve pour sa sang corp lumier larm
Stille Volk - Corps Magicien	de les ell dan la se le du pass est sa par mond leur corp vivr lair voyag feu
Florent Pagny - Tue-Moi	si plus que un tu mon mes jour souvenir parc

Capítulo 4

Funciones de divergencia

En el capítulo anterior se expuso la factorización de matrices no negativas y se abordó la necesidad de escoger una función de costo o pérdida adecuada. En este capítulo se describe la familia de divergencias que es más frecuentemente usada para la FMN.

La elección de la función divergencia define la forma de cuantificar las diferencias entre las matrices involucradas. Esto implica asumir ciertas propiedades estadísticas en los datos. En el capítulo 6 se exploran las consecuencias de diferentes funciones de costo y se relaciona la elección de cierta función con un modelo estadístico correspondiente.

Para la función de divergencia se puede considerar cualquier función que cumpla las siguientes características:

- $d : \mathbb{X}^2 \to \mathbb{R}$
- $d(u, v) \geq 0 \qquad \forall\ u,\ v \in \mathbb{X}$
- $d(u, v) = 0 \iff u = v \qquad \forall\ u,\ v \in \mathbb{X}$

Cabe destacar que a diferencia de las distancias, las divergencias no necesitan cumplir simetría ni la desigualdad del triángulo.

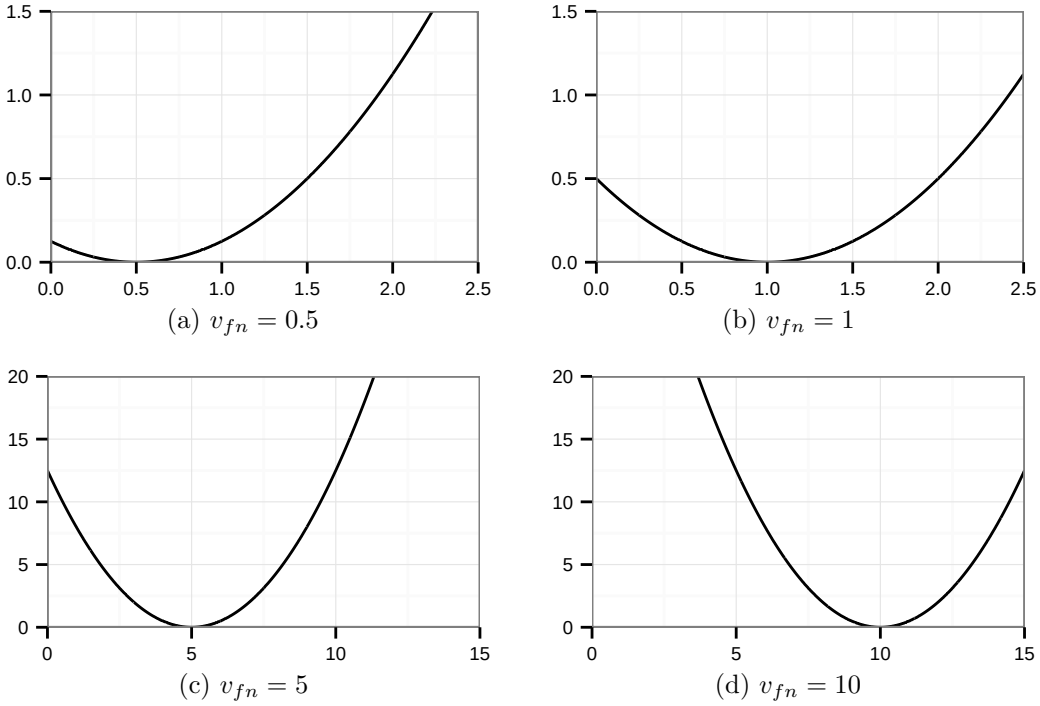
4.1. Casos Especiales

Aquí se presentan algunas de las divergencias más comunmente utilizados en la factorización de matrices no negativas.

4.1.1. Divergencia Euclideana

La divergencia euclidiana es probablemente la más usada en la mayoría de las aplicaciones. Su origen está en la forma tradicional de medir distancias en el plano, y la forma de calcularla es:

$$d_{\mathrm{EUC}}\left(v_{fn} \mid \hat{v}_{fn}\right) \quad = \quad \frac{1}{2}(v_{fn} - \hat{v}_{fn})^2$$



(a) $v_{fn} = 0.5$ (b) $v_{fn} = 1$

(c) $v_{fn} = 5$ (d) $v_{fn} = 10$

Figura 4.1: Divergencia Euclidiana $d_{\mathrm{EUC}}\left(v_{fn} \mid \hat{v}_{fn}\right)$ como función de $\hat{v}_{fn}$.

4.1.2. Divergencia Kullback - Leibler

Esta divergencia fue presentada por Kullback y Leibler [25] y se ha usado extensivamente para medir la diferencia entre dos distribuciones de probabilidad; en específico, sirve para medir la cantidad de información que se pierde por aproximar una distribución con otra. Algunos ejemplos de usos de esta divergencia incluyen la comparación de cadenas de Markov Rached et al. [34], la aproximación de funciones de densidad Georgiou y Lindquist [17], la extracción automática de texturas de

imágenes Do y Vetterli [9] y la robustificación de las inferencias en estudios ecológicos Burhnam y Anderson [4].

$$d_{\mathrm{KL}}\left(v_{fn} \mid \hat{v}_{fn}\right) \quad = \quad v_{fn} \log \frac{v_{fn}}{\hat{v}_{fn}} - v_{fn} + \hat{v}_{fn}$$

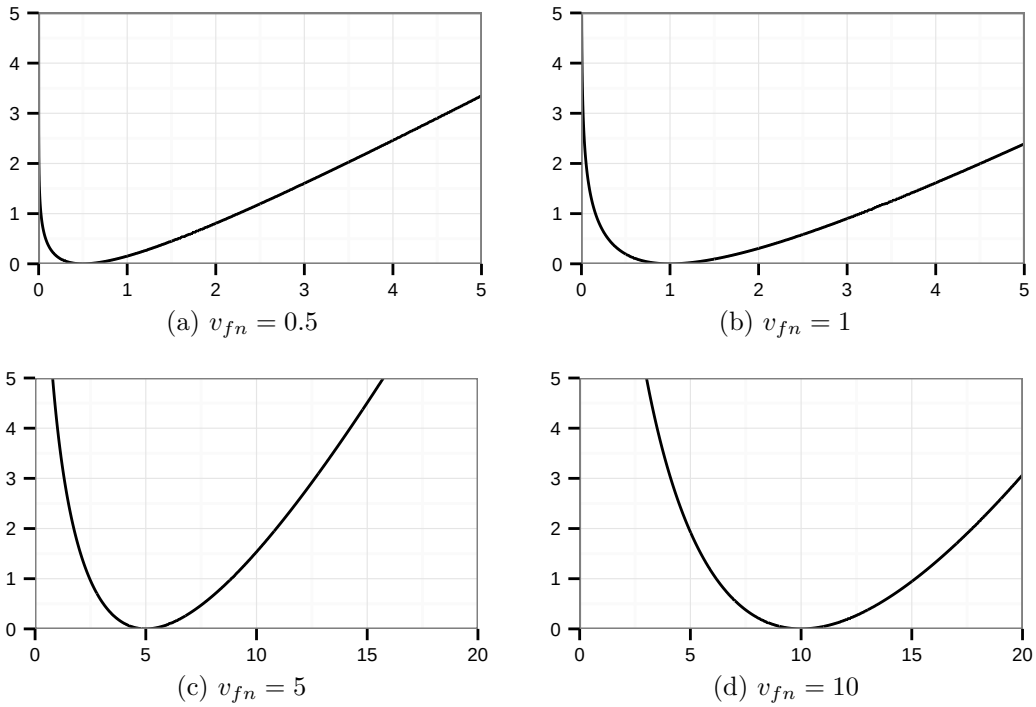


Figura 4.2: Divergencia de Kullback-Leibler $d_{\mathrm{KL}}\left(v_{fn} \mid \hat{v}_{fn}\right)$ como función de $\hat{v}_{fn}$.

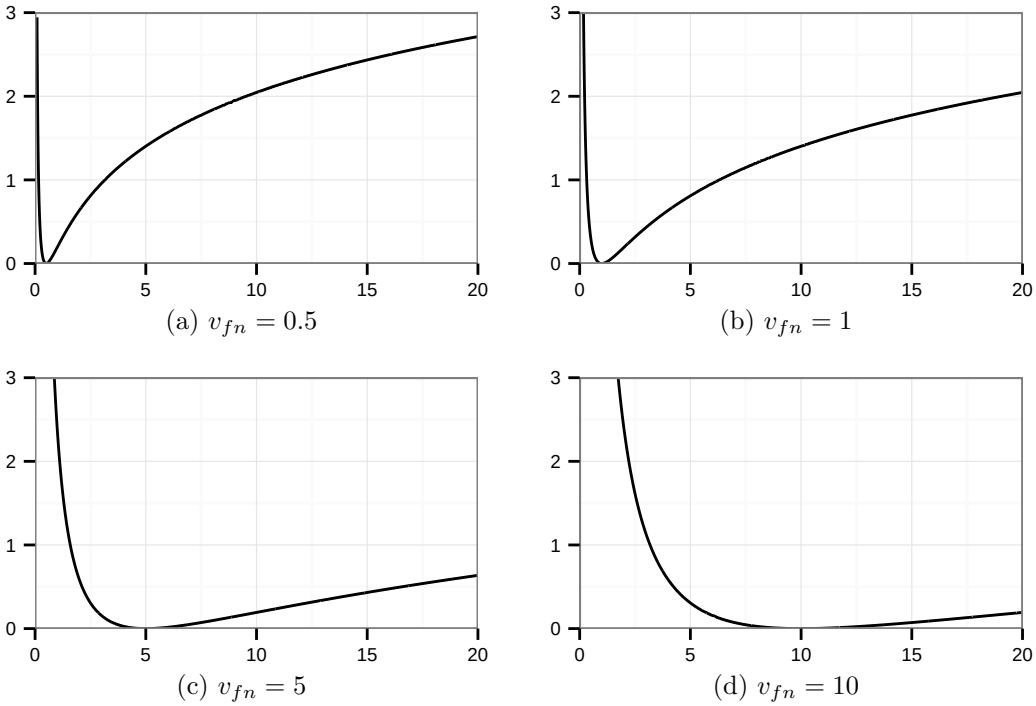
Las divergencias de Kullback-Leibler y de Euclides comparten la propiedad de que son convexas como funciones $\hat{v}_{fn}$). Esta característica se emplea en las demostraciones de convergencia de los algortimos de optimización numérica del apéndice A.

4.1.3. Divergencia Itakura-Saito

La divergencia Itakura-Saito se obtuvo por primera vez por Itakura y Saito [20] al maximizar la verosimilitud del espectro de audio usando un modelo autoregresivo y fue presentada como una forma de medir la bondad de ajuste entre dos espectros de audio. Esta divergencia ha sido usada recientemente para el análisis de espectogramas

de audio usando FMN; este trabajo ha sido desarrollado por Févotte et al. [14].

$$d_{\mathrm{IS}}\left(v_{fn} \mid \hat{v}_{fn}\right) \quad = \quad \frac{v_{fn}}{\hat{v}_{fn}} - \log \frac{v_{fn}}{\hat{v}_{fn}} - 1$$



(a) $v_{fn} = 0.5$ (b) $v_{fn} = 1$

(c) $v_{fn} = 5$ (d) $v_{fn} = 10$

Figura 4.3: Divergencia de Itakura-Saito $d_{\mathrm{IS}}\left(v_{fn} \mid \hat{v}_{fn}\right)$ como función de $\hat{v}_{fn}$.

Una propiedad interesante de la divergencia Itakura-Saito es que es invariante ante cambios de escala, esto es:

$$d_{\mathrm{IS}}\left(\lambda v_{fn} \mid \lambda \hat{v}_{fn}\right) \quad = \quad d_{\mathrm{IS}}\left(v_{fn} \mid \hat{v}_{fn}\right) \qquad \forall \lambda > 0$$

Esta propiedad significa que se considera con igual importancia relativa los coeficientes chicos y grandes de V en la función de costo, en el sentido de que un mal ajuste para un coeficiente pequeño estará igualmente penalizado que un mal ajuste para un coeficiente grande. Esto es diferente a las divergencias euclideana y Kullback-Leibler que asignan una mayor pérdida a las divergencias en los coeficientes grandes.

Por otro lado, a diferencia de las otras funciones presentadas anteriormente en este capítulo, $d_{\mathrm{IS}}(v_{fn} \mid \hat{v}_{fn})$ no es una función convexa como función de $\hat{v}_{fn}$.

4.2. β-Divergencia

La familia de β-divergencias es una familia parametrizadas por un parámetro β que contiene a las divergencias Euclidiana, Kullback-Leibler e Itakura-Saito como casos especiales ($\beta = 2, 1, 0$ respectivamente). La β-divergencia fue presentada por primera vez por Basu et al. [1]. Una discusión detallada sobre las propiedades de esta familia se puede encontrar en Cichocki y Amari [6].

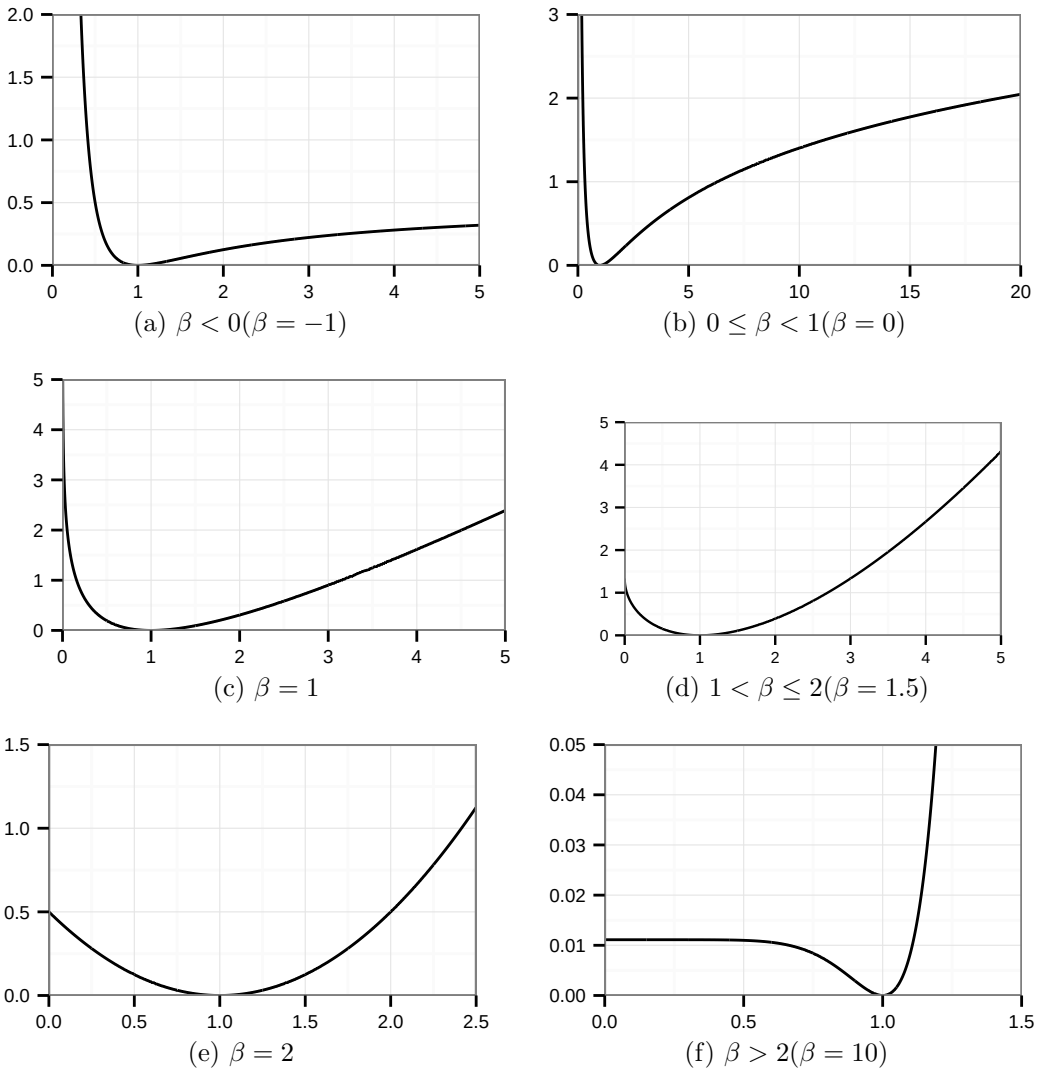
$$\begin{aligned} d_{\beta}(v_{fn} \mid \hat{v}_{fn}) &= \frac{1}{\beta(\beta-1)} v_{fn}^{\beta} + (\beta-1)\hat{v}_{fn}^{\beta} - \beta v_{fn}\hat{v}_{fn}^{\beta-1} \\ &= \frac{v_{fn}^{\beta} - v_{fn}\hat{v}_{fn}^{\beta-1}}{\beta-1} - \frac{v_{fn}^{\beta}}{\beta} + \frac{\hat{v}_{fn}^{\beta}}{\beta} \end{aligned}$$

Se puede mostrar fácilmente que:

$$\begin{aligned} d_{\mathrm{EUC}}(v_{fn} \mid \hat{v}_{fn}) &= d_{\beta}(v_{fn} \mid \hat{v}_{fn})|_{\beta=2} \\ d_{\mathrm{KL}}(v_{fn} \mid \hat{v}_{fn}) &= \lim_{\beta\to 1} d_{\beta}(v_{fn} \mid \hat{v}_{fn}) \\ d_{\mathrm{IS}}(v_{fn} \mid \hat{v}_{fn}) &= \lim_{\beta\to 0} d_{\beta}(v_{fn} \mid \hat{v}_{fn}) \end{aligned}$$

usando la identidad $\lim_{\beta\to 0} \frac{x^{\beta}-y^{\beta}}{\beta} = \log\frac{x}{y}$ para calcular los límites.

Existen otras familias de divergencias diferentes a las que se presentan en este capítulo; por ejemplo, las divergencias de Bregman (que son una generalización de la familia de β-divergencias) y las divergencias de Csiszár. Estas familias también han sido usadas para FMN por Dhillon y Sra [8] y Cichocki et al. [7] respectivamente.



(a) $\beta < 0 (\beta = -1)$ (b) $0 \leq \beta < 1 (\beta = 0)$

(c) $\beta = 1$ (d) $1 < \beta \leq 2 (\beta = 1.5)$

(e) $\beta = 2$ (f) $\beta > 2 (\beta = 10)$

Figura 4.4: β-divergencia $d_\beta \left(v_{fn} \mid \hat{v}_{fn} \right)$ como función de $\hat{v}_{fn}$ (con $v_{fn} = 1$) para diferentes valores de β.

Capítulo 5

Familia de Distribuciones Tweedie

En este capítulo se presenta la familia de distribuciones Tweedie que se usará para construir el modelo estadístico que servirá de base para la elección de la función de pérdidas en el problema de factorización de matrices no negativas. La familia de distribuciones Tweedie tiene como casos especiales a las distribuciones Normal, Gamma y Poisson y es un caso especial de los modelos de dispersión exponencial.

5.1. Modelos de Dispersión Exponencial

Los modelos de dispersión exponencial (MDE) son una familia de distribuciones con dos parámetros que consiste en una familia lineal exponencial con un parámetro de dispersión adicional. Esta familia de modelos es importante en estadística por su uso en los modelos lineales generalizados y fueron establecidos como un campo de estudio de interés por Jørgensen [23] que estudió con detalle sus propiedades.

Definición (Modelo de Dispersión Exponencial)**.** Sea Y una variable aleatoria cuya función de distribución se puede escribir de la forma:

$$f(y;\theta,\phi) = a(y,\phi)\exp\left[\frac{1}{\phi}(y\theta - \kappa(\theta))\right]$$

Entonces se dice que Y es un modelo de dispersión exponencial con parámetro de dispersión $\phi > 0$ y parámetro canónico θ.

Cualquier MDE se puede caracterizar por su función varianza $V()$, que describe la relación entre la media y la varianza de la distribución cuando se fija el parámetro de dispersión. Si Y sigue la distribución de un MDE con media μ, función varianza

$V()$ y parámetro de dispersión ϕ, entonces la varianza de Y se calcula como:

$$var(Y) = \phi V(\mu)$$

en donde ϕ es el parámetro de dispersión.

$\kappa(\theta)$ es llamada la función cumulante. Si $\phi = 1$, la media y varianza de la distribución se pueden calcular como $\kappa'(\theta) = \mu$ y $\kappa''(\theta) = V(\mu)$ respectivamente. En general, no se puede obtener una expresión analítica para $a(y, \phi)$, lo cual resulta en que sea difícil de evaluar la función de distribución.

Los modelos de dispersión exponencial tienen una función generadora de momentos simple, que se suele usar para evaluarlos. La función generadora de momentos es $M(t) = \int exp(ty) f(y; \theta, \phi) dy$. Sustituyendo la función de densidad, se obtiene que la función generadora de momentos es:

$$M(t) = exp\left(\frac{\kappa(\theta + t\phi) - \kappa(\theta)}{\phi}\right)$$

5.2. Distribuciones Tweedie

Definición (Familia de distribuciones Tweedie)**.** Sea Y un modelo de dispersión exponencial con media μ y parámetro de dispersion ϕ tal que la varianza cumple $V(\mu) = \phi\mu^p,\ p \in \Re\backslash(0, 1)$. Entonces se dice que Y es una distribución de la familia Tweedie con parámetro p.

La familia de distribuciones Tweedie incluye la mayoría de las distribuciones comunmente asociadas con los modelos lineales generalizados, incluyendo la distribución normal ($p = 0$), Poisson ($p = 1$), gamma ($p = 2$), entre otras. A pesar de que otros modelos Tweedie han sido menos estudiados y son menos conocidos, esta familia de distribuciones está bien definida para todos los valores de $p \in \Re\backslash(0, 1)$.

Cuando $p \geq 1$, todas las distribuciones de la familia Tweedie tienen dominio no negativo y medias estrictamente positivas, $\mu > 0$. Por otro lado, las distribuciones de la familia Tweedie en las que $p < 0$ son inusuales ya que cuentan con soporte en toda la recta real pero media estrictamente positiva. El objetivo de este trabajo es usar las distribuciones Tweedie como modelo estadístico en términos del cual se plantea un problema de inferencia que sea equivalente a la FMN. Debido a esto, las

distribuciones con $p < 0$ no son de interés puesto que no satisfacen la restricción de no negatividad de su dominio y no serán consideradas más en este trabajo.

Las distribuciones de la familia Tweedie son particularmente favorables para modelar datos continuos y positivos puesto que son los únicos MDE cerrados bajo reescalamiento de las variables. Las distribuciones en las que $1 < p < 2$ son particularmente atractivas cuando existen ceros exactos en los datos debido a que son distribuciones con masa de probabilidad en cero y densidad continua en los números positivos. Dunn y Smyth [12] señalan algunas de las aplicaciones en las que esta familia de distribuciones es usada e incluye estudios actuariales, de supervivencia, de series de tiempo, de consumo y meteorología.

La función cumulante y la media (parámetro canónico) que satisfacen la propiedad que caracteriza a las distribuciones de la familia Tweedie, $V(\mu) = \mu^p$, son:

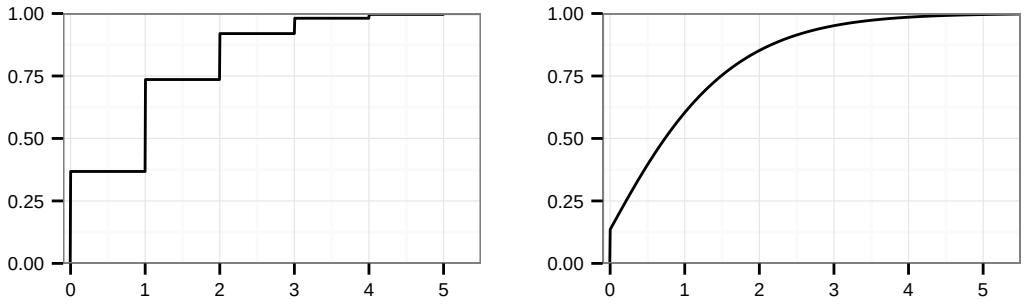
$$\theta = \begin{cases} \frac{\mu^{1-p}}{1-p}, & \text{si } p \neq 1 \\ \log \mu, & \text{si } p = 1 \end{cases}, \qquad \kappa(\theta) = \begin{cases} \frac{\mu^{2-p}}{2-p}, & \text{si } p \neq 2 \\ \log \mu, & \text{si } p = 2 \end{cases}.$$

Desafortunadamente, sólo las distribuciones con $p = 0, 1, 2$ o 3 cuentan con una función de densidad analítica lo cual limita el uso de estas distribuciones en la práctica. Por el otro lado, las distribuciones de la familia Tweedie cuentan con una función generadora de momentos simple, lo que se usa como base para evaluar numericamente estas distribuciones.

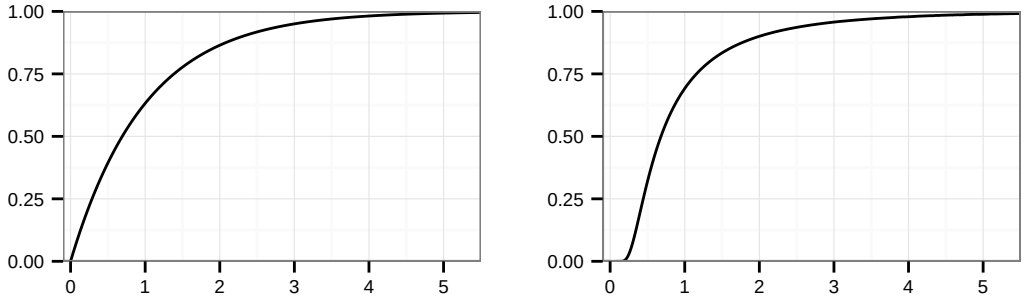
Dunn y Smyth [11] exploran alternativas para evaluar las distribuciones usando expansiones de series que aproximan la funcion de densidad. Estos trabajos son complementados por Dunn y Smyth [13], que presentan una alternativa usando análisis de Fourier. Dunn [10] presenta la implementación en el software estadístico R de los métodos de evaluación que se presentan en este trabajo y que fue utilizado para producir los resultados de este trabajo.

5.2.1. Casos Especiales

Es fácil ver que las distribuciones Normal, Gamma y Poisson corresponden a densidades de la familia Tweedie con parámetros 0, 1 y 2 respectivamente. Estas distribuciones son bien conocidas en la literatura. Para una mejor referencia sobre sus propiedades se puede consultar Krishnamoorthy [24].



(a) $p = 1$ (b) $1 \leq p < 2 \quad (p = 1.5)$



(c) $p = 2$ (d) $2 < p \quad (p = 4)$

Figura 5.1: Función de probabilidad acumulada para distribuciones de la familia Tweedie con distintos valores de p (con $\phi = 1$ y $\mu = 1$).

Distribución	**p**	**Función de distribución**
Normal	0	$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{(\frac{x-\mu}{\sigma})^2}$
Poisson	1	$f(x;\lambda) = \frac{\lambda^x}{x!} e^{-\lambda} I_{0,1,\ldots}(x)$
Gamma	2	$f(x;a,b) = \frac{x^{a-1} e^{-bx} b^a}{\Gamma(\alpha)}$

Cuadro 5.1: Casos especiales de distribuciones de la familia Tweedie

Capítulo 6

Modelo Estadístico para FMN

La interpretación de FMN como una aproximación de bajo rango, en el sentido de minimizar una función de distancia d, puede ser suficiente para la derivación de algortimos útiles para diferentes aplicaciones, como por ejemplo para descomponer señales. Sin embargo, para diversas aplicaciones no es claro cual es la función de distancia apropiada. La selección de la función de costo está relacionada con las propiedades estadísticas de V y se puede intentar resolver planteando un modelo estadístico desde el cual se pueda inferir el parámetro adecuado. La selección de una función de distancia d para medir la calidad del ajuste entre V y $\hat{V}$ implica asumir supuestos sobre la forma en que V es generada de $\hat{V}$. En la literatura se ha hecho referencia previamente a los modelos estadísticos equivalentes al problema de FMN.

$$v_{fn} \sim \mathcal{N}(v_{fn}; \hat{v}_{fn}, \sigma^2) \qquad \text{FMN} - \text{EUC}$$
$$v_{fn} \sim \mathcal{P}(v_{fn}; \hat{v}_{fn}) \qquad \text{FMN} - \text{KL}$$
$$v_{fn} \sim \mathcal{G}(v_{fn}; a, \frac{a}{\hat{v}_{fn}}) \qquad \text{FMN} - \text{IS}$$

Un detalle que es muy importante resaltar es que en el caso del modelo estadístico la maximización de la verosimilitud se realiza en simultáneo para todos los parámetros del modelo. Este hecho, aunado a las restriccciones de no negatividad, es lo que permite afrontar el problema de no identificación.

6.1. Modelos Compuestos

Siguiendo el trabajo de Févotte y Cemgil [15], aquí se hace uso de la propiedad de cerradura bajo suma de las distribuciones Gaussiana y Poisson; cuando $x = \sum_k c_k$ y c_k se distribuye Gaussiana (o Poisson), x se distribuye Gaussiana (o Poisson). Se considera el modelo generador:

$$\begin{aligned} v_{fn} &= \sum_k c_k \\ c_{k,fn} &\sim p(c_{k,fn}|\theta_k) \end{aligned}$$

donde $\theta_k = \{w_{f,k}\}_{f=1}^{F} \cup \{w_{k,n}\}_{n=1}^{N}$. A continuación se expone a detalle el modelo estadístico equivalente a la factorización de matrices no negativa con las divergencias euclidiana, Kullback-Leibler e Itakura-Saito.

6.1.1. FMN con la Divergencia Euclideana

Se supone el siguiente modelo generador:

$$c_{k,fn} \sim \mathcal{N}(c_{k,fn}; w_{fk}h_{kn}, \frac{\sigma^2}{K})$$

entonces se puede probar que:

$$-\log p(V|W,H,\sigma^2) = \frac{1}{\sigma^2} D_{\text{EUC}}(V|WH) + \frac{NF}{2}\log(2\pi\sigma^2)$$

Así pues, la estimación de máxima verosimilitud de W y H es equivalente a resolver el problema de FMN usando la divergencia Euclidiana.

Existe, sin embargo, ambigüedad en la interpretación del modelo generado por las ecuaciones, ya que podría producir datos negativos. Como tal, aun cuando el problema de optimización resultante es el mismo siempre que los datos en V sean no negativos. Se podría usar la distribución normal truncada, pero esto rompería la correspondencia debido al poceso de renormalización de la distribución.

6.1.2. FMN con la Divergencia Kullback-Leibler

Se supone el siguiente modelo generador:

$$c_{k,fn} \quad \sim \quad \mathcal{P}(c_{k,fn}; w_{fk}h_{kn})$$

entonces se puede probar que:

$$-\log p(V|W,H) \quad \stackrel{c}{=} \quad D_{\mathrm{KL}}(V|WH)$$

donde $\stackrel{c}{=}$ denota igualdad salvo por una constante. Esto quiere decir que la estimación de máxima verosimilitud de este modelo es equivalente a resolver FMN usando la divergencia KL. Aun cuando los datos generados bajo este modelo estadstico son no negativos, existe ambigüedad en la interpretación del modelo debido a que la distribución Poisson sólo considera en su dominio números enteros.

6.1.3. FMN con la Divergencia Itakura-Saito

Como Févotte et al. [14] describe, existe una equivalencia entre la factorización de matrices no negativas usando la divergencia de Itakura-Saito y el modelo estadístico que supone que:

$$V \quad = \quad (WH).E$$

donde E es una matriz en la que cada entrada se distribuye gamma con media 1. Si se supone este modelo estadístico, entonces se puede probar que:

$$-\log p(V|W,H) \quad \stackrel{c}{=} \quad D_{\mathrm{IS}}(V|WH)$$

Por lo tanto, la estimación de máxima verosimilitud de W y H en dicho modelo estadístico es equivalente a resolver el problema de FMN usando la divergencia Itakura-Saito.

En el capítulo 7 se hace uso de los modelos estadísticos expuestos en esta sección para plantear un modelo estadístico general que permita la selección automática de β en la FMN usando la β-divergencia.

Capítulo 7

Selección de β

La idea detrás del algortimo que se plantea en esta sección es hacer uso de la equivalencia entre el problema estadístico de maximización de verosimilitud y el problema de factorización dematrices para tener un mecanismo para elegir la función de costos.

7.1. Relación entre la elección de la divergencia y distribuciones Tweedie

Lo más importante a destacar es que los casos especiales estudiados sugieren que existe una relación directa entre el parámetro β de la divergencia que se usa para FMN y del parámetro p de la distribución Tweedie usada en el modelo generador. La relación es que al plantear un modelo estadístico generador con parámetro p, esto es equivalente a minimizar la divergencia con parámetro $\beta = 2 - p$ en los casos especiales que se consideran en este trabajo. Partiendo de este hecho y siguiendo un principio de parsimonia, se supone que la relación $\beta = 2 - p$ se mantiene fuera de los casos considerados. Este supuesto es fundamental en el planteamiento del problema y desafortunadamente no se cuenta con una demostración en este trabajo. La dificultad de esta prueba radica en la falta de una expresión cerrada para las distribuciones Tweedie en general.

Partiendo de este hecho, se propone el algoritmo de selección de β para factorización de matrices no negativas usando la β divergencia.

7.2. Selección de β en FMN usando la β divergencia.

La idea principal detrás del algoritmo que se presenta a continuación es que existe una relación directa entre el parámetros p del modelo estadístico equivalente y el algoritmo de factorización de matrices no negativas usando la β-divergencia. Como se explicó anteriormente, se supone que la relación se cumple cuando el modelo estadístico se plantea con distribuciones de la familia Tweedie y corresponde a la relación entre parametros $\beta = 2 - p$. El problema de selección del modelo estadístico adecuado se ha tratado con cuidado en la literatura; por ejemplo por Burnham y Anderson [5].

El algoritmo se basa en separar el problema en 2 fases. En la primera fase, se aproxima la factorización de matrices usando las reglas multiplicativas explicadas en el apéndice A fijando β. En la segunda fase, se aproxima el parámetro p de la familia Tweedie óptimo de acuerdo a la verosimilitud en un modelo estadístico, fijando las matrices W y H. Estas dos etapas se repiten alternativamente hasta encontrar convergencia numérica del algoritmo.

Inicializar aleatoriamente β, W y H;
mientras $D_\beta(V||\hat{V}) > TOL.$ **hacer**
  para $i = 1, ..., n_1$ **hacer**

$$W \leftarrow W.\frac{[V.\hat{V}^{\beta-2}]H^T}{\hat{V}^{\beta-1}H^T}, \qquad H \leftarrow H.\frac{W^T[V.\hat{V}^{\beta-2}]}{W^T\hat{V}^{\beta-1}}$$

  fin
  para $i = 1, ..., n_2$ **hacer**

$$p = p + \alpha\frac{\partial \log p(p; \hat{V}, \phi)}{\partial p}$$

  fin
  $\beta = 2 - p$
fin

Algoritmo 1: Selección de β.

7.3. Resultados

La implementación de este algoritmo tiene como parámetros $n_1 = 1000$, $n_2 = 100$, $TOL = 10^{-6}$ y $\alpha = .1$, cabe destacar que la derivada de la derivada de la log-verosimilitud se obtiene con una aproximación numérica. Esta implementación se aplicó a datos del espectrograma que se presenta en la sección 1.5.3 se encontró que el parámetro se debería usar la β divergencia con parámetro 0.012862 después de 100000 iteraciones del algoritmo planteado en la sección anterior. Vale la pena notar que como se mencionó anteriormente, es muy común en la literatura usar la divergencia Itakura-Saito en los datos provenientes de espectogramas, como se describe en Févotte et al. [14]. Así, parece que al menos este resultado es consistente con la literatura anterior.

7.4. Conclusiones

Este trabajo aborda un problema específico en el área de aprendizaje estadístico no supervisado, la selección de la función de divergencia en la factorización de matrices no negativas. La importancia de este trabajo consiste en simplificar la elección de la función de pérdidas sujetándola a características de los datos.

Existen diferentes posibilidades para evaluar este algoritmo que no fueron desarrolladas en este trabajo y que permitirían demostrar el alcance o las limitaciones del algoritmo. La primera opción consiste en aplicar el algoritmo a distintas bases de datos de carácter musical en las cuales se busque recuperar parámetros parecidos para la función de costo. Por otro lado, es importante mencionar que los resultados de este trabajo dependen crucialmente de que la relación que existe entre p y β sea correctamente identificada. En este trabajo se supone que la relación es lineal, lo cual se cumple en los casos especiales en los que existe una distribución de forma cerrada y que se pueden evaluar exactamente.

Finalmente, hay dos lecciones valiosas que se pueden extraer del desarrollo de este trabajo. En primer lugar, la función de costos se puede extraer en términos de los datos del problema. En segundo lugar, la selección de la función de costos es equivalente a seleccionar el modelo estadístico que mejor ajusta los datos. Por lo tanto, se concluye que al escoger una función de costos en el problema de optimización, se está haciendo implícitamente un supuesto sobre las propiedades estadísticas de los

datos.

Apéndice A

Algoritmos de Solución

En este apéndice se explican los diferentes algoritmos de solución para el problema de factorización de matrices no negativas usando la β-divergencia.

Lee y Seung [26] presentan la primera versión del algoritmo para resolver la factorización de matrices no negativas usando la divergencia Kullback-Leibler. Este algoritmo consiste en empezar de condiciones no negativas escogidas aleatoriamente y actualizar las matrices de acuerdo a la siguiente regla de actualización:

$$W_{fk} \leftarrow W_{fk} \frac{\sum_n \frac{H_{kn} V_{nk}}{(WH)_{fn}}}{\sum_n W_{kn}}, \qquad H_{kn} \leftarrow W_{kn} \frac{\sum_f \frac{W_{fk} V_{nk}}{(WH)_{fn}}}{\sum_n H_{kn}}$$

La simplicidad y facilidad de implementar de este algoritmo resultó en gran popularidad y en su uso generalizado. Posteriormente, Lee y Seung [27] expanden y detallan los algoritmos antes presentados, incluyendo una versión parecida para la divergencia Euclidiana. En este mismo trabajo, se demuestra que al actualizar las matrices de esta forma, se logra una actualización no creciente y que solamente es invariante si WH es un puto estacionario de la divergencia. Cabe la pena destacar que cuando $V = WH$ esta actualización no cambia las matrices, con lo que uno de los puntos estacionarios de la divergencia es cuando se encuentra una reconstrucción completa.

En la literatura se pueden encontrar distintos esfuerzos para encontrar algoritmos más robustos o más eficientes que permitan extender el trabajo de Lee y Seung [26]. Así, se ha encontrado que la factorización de matrices, i.e., $\hat{V} = WH$, se puede

realizar de forma eficiente con las siguientes reglas multiplicativas de actualización.

$$W \leftarrow W.\frac{[V.\hat{V}^{\beta-2}]H^T}{\hat{V}^{\beta-1}H^T}, \qquad H \leftarrow H.\frac{W^T[V.\hat{V}^{\beta-2}]}{W^T\hat{V}^{\beta-1}}$$

que garantizan un descenso monótono de $D_\beta(V||\hat{V})$ cuando $\beta \in [1, 2]$. Una discusión detallada de las propiedades de convergencia de este algoritmo de optimización puede ser encontrada en por Févotte y Idier [16]. Las técnicas de optimización numérica que se usan en este problema en específico, se pueden encontrar en Nocedal y Wright [31].

También se han desarrollado nuevos algoritmos basados en la transformada rápida de Fourier que consiguen resultados similares en menor tiempo. Este trabajo, fue realizado principalmente por Li et al. [28].

Apéndice B

Software y Reproducibilidad

Una copia digital de este trabajo puede encontrarse en la dirección: https://bitbucket.org/jimsotelo/tesis.git. En esta dirección se encontraran todos los archivos necesarios para generar las figuras que este trabajo incluye, asi como el documento escrito y el código para replicar todos los resultados de este trabajo.

Software usado:

- make
- R
- LaTeX
- Python

Bibliografía

[1] Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, y M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, y Paul Lamere. The million song dataset. En *Proceedings of the 12th International Conference on Music Information Retrieval*. 2011.

[3] L. Breiman, J. Friedman, R. Olshen, y C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[4] Kenneth Burhnam y David Anderson. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, 28(2):111–119, 2001.

[5] K.P. Burnham y D.R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, 2002.

[6] Andrzej Cichocki y Shun-ichi Amari. Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[7] Andrzej Cichocki, Rafal Zdunek, y Shun-ichi Amari. Csiszrs divergences for non-negative matrix factorization: Family of new algorithms. En Justinian Rosca, Deniz Erdogmus, José C. Príncipe, y Simon Haykin, eds., *Independent Component Analysis and Blind Signal Separation*, tomo 3889 de *Lecture Notes in Computer Science*, págs. 32–39. 2006.

[8] Inderjit S. Dhillon y Suvrit Sra. Generalized nonnegative matrix approximations with bregman divergences. En *NIPS*. 2005.

[9] M.N. Do y M. Vetterli. Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance. *IEEE Transactions on Image Processing*, 11(2):146–158, 2002.

[10] Peter K Dunn. *tweedie: Tweedie exponential family models*, 2014. R package version 2.2.1.

[11] Peter K. Dunn y Gordon K. Smyth. Tweedie family densities: Methods of evaluation. En *Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling*. 2001.

[12] Peter K. Dunn y Gordon K. Smyth. Series evaluation of tweedie exponential dispersion model densities. *Statistics and Computing*, 15:267–280, 2005.

[13] Peter K. Dunn y Gordon K. Smyth. Evaluation of tweedie exponential dispersion model densities by fourier inversion. *Statistics and Computing*, 18:73–86, 2008.

[14] Cédric Févotte, Nancy Bertin, y Jean-Louis Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[15] Cédric Févotte y A. Taylan Cemgil. Nonnegative matrix factorizations as probabilistic inference in composite models. En *17th European Signal Processing Conference*. 2009.

[16] Cédric Févotte y Jérome Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

[17] T.T. Georgiou y A Lindquist. Kullback-leibler approximation of spectral density functions. *IEEE Transactions on Information Theory*, 49(11):2910–2917, 2003.

[18] Trevor Hastie, Robert Tibshirani, y Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[19] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 1933.

[20] Fumitada Itakura y Shuzo Saito. Analysis synthesis telephony based on the maximum likelihood method. En *Proceedings of the 6th International Congress on Acoustics*. 1968.

[21] Gareth James, Daniela Witten, Trevor Hastie, y Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

[22] Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2004.

[23] Bent Jørgensen. Exponential dispersion models. *Journal of the Royal Statistical Society*, 49:127–162, 1987.

[24] K. Krishnamoorthy. *Handbook of Statistical Distributions with Applications*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 2006.

[25] Solomon Kullback y Richard Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951.

[26] Daniel D. Lee y H. Sebastian Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[27] Daniel D. Lee y H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001.

[28] Liangda Li, Guy Lebanon, y Haesun Park. Fast bregman divergence nmf using taylor expansion and coordinate descent. En *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, págs. 307–315. ACM, 2012.

[29] Stanislav Nikolov. Principal component analysis : Review and extensions. 2010.

[30] Nils J. Nilsson. *Introduction to Machine Learning: An Early Draft of a Proposed Textbook*. 1996.

[31] Jorge Nocedal y Stephen J. Wright. *Numerical Optimization*. Springer, 2006.

[32] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

[34] Z. Rached, F. Alajaji, y L.L. Campbell. The kullback-leibler divergence rate between markov sources. *Information Theory, IEEE Transactions on*, 50(5):917–921, 2004.

[35] Stuart Russell y Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, 2009.

[36] Wenwu Wang. *Machine audition : principles, algorithms, and systems*. Information Science Reference, 2011.