UNIVERSIDAD CENTROCCIDENTAL
"LISANDRO ALVARADO"

Decanato de Ciencias y Tecnología
Licenciatura en Ciencias Matemáticas

"Problemas Inversos en Estadística no Paramétrica"

Trabajo Especial de Grado presentado por

Br. Jhon A. Rodríguez P.

como requisito final
para obtener el título de Licenciado
en Ciencias Matemáticas
Área de Conocimiento: **Matemática Aplicada**
Tutor: Dr. Hugo Lara Urdaneta

Barquisimeto, Venezuela. Junio de 2009

Universidad Centroccidental
"Lisandro Alvarado"
Decanato de Ciencias y Tecnología
Licenciatura en Ciencias Matemáticas

ACTA
TRABAJO ESPECIAL DE GRADO

Los suscritos miembros del Jurado designados por el Jefe del Departamento de Matemáticas del Decanato de Ciencias y Tecnología de la Universidad Centroccidental "Lisandro Alvarado", para examinar y dictar el veredicto sobre el Trabajo Especial de Grado titulado:

"Problemas Inversos en Estadística no Paramétrica"

Presentado por el ciudadano Br. Jhon A. Rodríguez P. titular de la Cédula de Identidad Nº V-16.749.348. Con el propósito de cumplir con el requisito académico final para el otorgamiento del título de Licenciado en Ciencias Matemáticas.

Luego de realizada la Defensa y en los términos que imponen los Lineamientos para el Trabajo Especial de Grado de la Licenciatura en Ciencias Matemáticas, se procedió a discutirlo con el interesado habiéndose emitido el veredicto que a continuación se expresa:

[1] _____

Con una calificación de _____ puntos.

En fe de lo expuesto firmamos la presente Acta en la Ciudad de Barquisimeto a los _____ días del mes de _____ de _____.

TUTOR	FIRMA
JURADO	FIRMA
JURADO	FIRMA

OBSERVACIONES:

[1] Aprobado ó Reprobado

Universidad Centroccidental "Lisandro Alvarado"

"Problemas Inversos en Estadística no Paramétrica"

Br. Jhon A. Rodríguez P.

Tutor: Dr. Hugo Lara Urdaneta

Resumen

Los problemas inversos, donde deseamos encontrar ciertas cantidades que no son directamente medibles, surgen en diversas áreas en ciencias e ingeniería. Cuando dichos problemas son mal puestos o mal condicionados se dificulta la reconstrucción de las cantidades deseadas. La teoría de regularización se presenta como alternativa para extraer dicha información de los problemas inversos, cuando están mal puestos o mal condicionados. El presente trabajo pretende estudiar la teoría de los problemas inversos, en el marco de la estadística no paramétrica, considerando el modelo de ruido blanco, donde el ruido presente es considerado aleatorio.

Palabras Clave: Problemas inversos, Métodos de Regularización, Estadística no paramétrica, ruido blanco.

Dedicado a mis padres Hector E. y Pastora del C. y a mis hermanas Laudys y Johana.

AGRADECIMIENTOS

A Dios Todopoderoso, mi Padre Celestial, por concederme el poder concluir este trabajo.

A mi familia por la confianza que depositan en mí.

A mis amigos, por el apoyo y el ánimo que me han brindado. A mis compañeros y amigos de promoción con quienes he compartido momentos de estudio y nos hemos ayudado para llegar hasta aquí.

A los demás compañeros de quienes a lo largo de la carrera me brindaron su ayuda.

A mis profesores por la formación académica que me han dado.

En fin, a todos quienes con su apoyo y estímulo me ayudaron a hacerlo posible. Gracias.

ÍNDICE

Índice de Figuras

INTRODUCCIÓN

Los problemas inversos se presentan en muchos estudios en el ámbito de la matemática aplicada, cuando se pretenden resolver diversos problemas provenientes de la ingeniería y de otras ciencias aplicadas. Básicamente, en estos campos la herramienta con que se cuenta al momento de realizar un estudio de algún fenómeno físico, por ejemplo, es la recolección de datos derivados de observaciones que se hacen al fenómeno. Siempre dichas observaciones serán aproximaciones de lo que realmente ocurre, ya que por distintas variables, estas no son exactas; hay que contar con el error que se comete al tomar los datos. Aquí se podría mencionar la diferencia en la percepción del fenómeno que podrían tener dos personas distintas que toman la muestra, o la calibración que tenga un instrumento usado para obtener los datos.

Si los datos medidos dependen, de alguna forma, de las cantidades que queremos, entonces estos contienen al menos cierta información acerca de dichas cantidades. Comenzando con los datos que hemos medido, el problema de tratar de reconstruir las cantidades que realmente queremos es un "problema inverso". De manera coloquial, decimos que un problema inverso es cuando medimos un efecto y queremos determinar su "causa". [6]

En la actualidad, son muchos los campos donde se plantean problemas inversos en pro de conocer más a fondo fenómenos, para satisfacer las demandas que van haciendo distintas necesidades presentes en la vida moderna. Así, podríamos citar ejemplos en áreas como la tomografía axial computarizada: que toma radiografías en varios ángulos y reconstruye la imagen tridimensional (representada en cortes transversales). Ajuste de modelos: donde se asume un modelo de los datos, y se determinan los parámetros que mejor ajustan los datos; problemas de deconvolución, donde una imagen fotográfica difusa representa los datos observados, y se desea obtener la imagen original; imagenes

radioastronómicas, navegación, análisis de imagenes (conteo de estrellas, o glóbulos rojos, por ejemplo); en geofísica, describiendo la estructura interior de la Tierra, etc. (Para más ejemplos ver [7])

Dado todo este panorama donde se encuentran presentes los problemas inversos, cada vez más crece el interés por realizar estudios en esta área, y así contribuir al desarrollo del conocimiento que se tiene al respecto, aportando soluciones óptimas a estas ciencias que así lo demandan.

En el presente trabajo se pretende realizar un estudio de los problemas inversos lineales, el mal condicionamiento y la mala postura de los mismos. Se plantea la regularización como alternativa en la solución de algunos problemas, y los inconvenientes que puede generar la presencia de un factor de ruido al momento de resolver un problema. En esta primera parte, se basó en una revisión realizada al material de Tan and Colin.[6]

Por otro lado, en el capítulo 2 se realiza una descripción del trabajo desarrollado en [1] por Cavalier, para los problemas inversos en estadística no paramétrica. Acá, se explican temas clásicos, como el modelo de ruido blanco, estimación de riesgo, riesgo minimax y otros temas de estudio más reciente, como estimación adaptada, desigualdades oráculos, para finalmente concluir con la estimación de riesgo insesgada de Stein y el método de la cápsula de riesgo. De este trabajo, se revisaron los resultados principales y los colocamos en el contexto de la teoría de los problemas inversos.

Capítulo 1

PRELIMINARES

§1.1. Breve introducción a los problemas inversos

En distintas áreas de las ciencias donde se propone la recolección de datos, se tiene como fin recabar información sobre algún fenómeno físico o de interés. Por supuesto, que en la mayoría de los casos los datos que se recolectan difieren de los que se quieren medir. Nos encontramos ante un problema inverso cuando intentamos reconstruir la data deseada partiendo de la data medida, esto es factible cuando los datos medidos dependen de alguna manera de los datos que se desean medir.
Citamos ahora algunos ejemplos de problemas inversos:

- **Modelo de Ajuste**
  La ecuación $y = a + bx + cx^2 + dx^3$ indica que el valor de "y"depende del valor que toma la variable "x". Si tenemos un conjunto de puntos $\{(x_i, y_i)\}_{i=1}^n$ el cual representa la data, cabe la pregunta ¿Cómo determinar los valores de a, b, c, y d que determinan el modelo dado?. En este caso la imágen que deseamos reconstruir son los valores entre a y d.

- **Deconvolución**
  Al tenerse una fotografía borrosa o una señal que ha atravesado algún medio, ¿Cómo reconstruir la imágen fotográfica no borrosa o la señal que se emitió originalmente?. Este tipo de problemas tiene aplicación directa en las comunicaciones, donde se envía una señal la cual se distorsiona al ser transmitida, por ejemplo, a través de un conductor (cable), haciendose necesario reconstruir la señal enviada originalmente.

- **Análisis Numérico**
  Solución de ecuaciones integrales tales como la ecuación de Fredholm de primera clase $\int_a^b k(x,s)z(s)ds = y(x)$, con $c \leqslant x \leqslant d$, donde el kernel k y la función y son

dadas, puede ser tratada como un problema inverso para la función z desconocida. El caso especial donde $k(x, s) = u(x - s)$ (siendo u la función escalón unidad) es el problema de diferenciación numérica de la función u.

§1.1.1. Algunas definiciones preliminares

Los ejemplos dados anteriormente constituyen una pequeña lista de la diversidad de áreas en las ciencias donde se pueden plantear problemas inversos. Ello sugiere que el estudio de estos es muy extenso. Una gran gama de ejemplos en otras muchas áreas es mencionada y descrita en [7]. Para comenzar este estudio necesitamos tener en cuenta algunas definiciones y así ir estableciendo la terminología que se usará a lo largo del trabajo.

La colección de valores que deseamos reconstruir la llamaremos **Imagen**; en el problema de deconvolución esta representa la fotografía que se quiere reconstruir, pero en los problemas inversos en general, serán parámetros que definen un modelo. La imagen la denotaremos por f. El conjunto de todas las imágenes lo definimos como el **espacio imagen**.

Estamos en presencia del **problema directo** cuando conocemos la aplicación que va del espacio imagen a los valores por nosotros medidos (data). La aplicación directa puede ser o no lineal y es denotada por A.

El conjunto de todos lo posibles datos es el **espacio de datos**. Los datos en un problema los denotaremos por d.

Denotando por $\overline{d}$ a la data libre de ruido, el problema directo se considera como la aplicación que va de la imágen a $\overline{d}$. Para nosotros el **ruido** será la diferencia $\overline{d} - d$ y lo denotaremos por n.

De todo esto podemos escribir la relación $d = A(f) + n$ como una aplicación directa, que va de la imagen más un ruido presente a la data actual.

Así, finalmente podemos definir el **problema inverso** como el problema de encontrar la imagen original dada la data y conociendo el problema (aplicación) directo.

Para el caso de deconvolución, donde se busca dar nitidez a una fotografía, la "imagen" es la fotografía clara, la "data" es la fotografía borrosa y el problema directo es el proceso de hacer borrosa la imagen original. El problema inverso consiste en encontrar la fotografía nítida (imagen) dada la fotografía borrosa (data) y conociendo el proceso que hace borrosa la imagen.

§1.1.2. Problemas mal puestos y mal condicionados

La imagen y la data pueden ser funciones de variable continua o discreta, de aquí se deriva una clasificación básica para el problema directo. Estos pueden ser: continuo-continuo, continuo-discreto, discreto-continuos o discretos-discretos. En la práctica los problemas se presentan del último tipo mencionado ya que los datos que se pueden medir siempre serán finitos y para la imagen se trabaja con una discretización de la misma. Los otros tipos de problemas siempre serán idealizaciones del problema original.

Sea cual sea el tipo de problema al cual nos enfrentamos, el problema inverso de resolver

$$A(f) = d \tag{1.1.1}$$

para "f"dado "d", se dirá **bien puesto** (Hadamard (1923)) si:

- Existe una solución para cualquier data en el espacio de datos.
- La solución es única en el espacio de imágenes.
- La aplicación inversa $d \to f$ es continua.

Las primeras dos condiciones son equivalentes a exigir que la inversa de A esté bien definida y que su dominio sea todo el espacio de datos.
La continuidad exigida en la tercera condición es una condición necesaria pero no suficiente para la estabilidad de la solución.

Cuando se presenta un problema bien puesto, el error de propagación relativo de la data a la solución es controlodo por el número de condición de A.

Si Δd es una variación de d y Δf la correspondiente variación de f, entonces:

$$\frac{\|\Delta f\|}{\|f\|} \leqslant cond(A)\frac{\|\Delta d\|}{\|d\|} \tag{1.1.2}$$

de allí que se desee un número de condición pequeño.

Si $cond(A)$ no es tan grande se dice que el problema

$$A(f) = d$$

está bien condicionado y la solución es estable con respecto a pequeñas variaciones de la data. En caso contrario se dice mal condicionado.

Hadamard definió un problema mal puesto a aquel que deja de satisfacer una de las tres condiciones dadas anteriormente. Así, se está al frente de un problema mal puesto cuando la inversa de A no existe porque la data no se encuentra en el rango de A o la imagen, que es solución del problema, no es única, ya que más de una imagen es enviada a la misma data, o porque un pequeño cambio arbitrario en la data produce un gran cambio arbitrario en la imagen; esta última es equivalente a no satisfacerse la tercera condición para que el problema sea bien puesto.

Hadamard pensó que los problemas mal puestos erán "artificiales", que no describían ningún fenómeno físico en particular. Estaba equivocado. Los ejemplos de problemas inversos citados al principio son todos mal puestos o al menos mal condicionados. El hecho que tomografías axiales computarizadas son realizadas a diario, o que reservas de crudo han sido encontradas por investigación sísmica es evidencia que información significativa puede ser ganada de problemas inversos mal puestos aún cuando ellos no pueden ser estrictamente invertidos.

§1.2. Estudio de una transformación lineal

§1.2.1. Autovalores y autovectores de una matriz simétrica

Consideremos M una matriz real simétrica de orden n. Sus autovalores son reales y sus autovectores forman una base ortonormal de $\mathbb{R}^n$. Denotando como μ_i a los autovalores y u_i el autovector asociado al autovalor μ_i se tiene:

$$Mu_i = \mu_i ui. \tag{1.2.1}$$

Para algunos resultados del Álgebra Lineal que se presentarán, se consultó [4]. Sea U la matriz cuyas columnas son los autovectores de M. Así,

$U = \begin{pmatrix} u_{11} & u_{21} & \cdots & u_{n1} \\ u_{12} & u_{22} & \cdots & u_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1n} & u_{2n} & \cdots & u_{nn} \end{pmatrix}$, donde u_{ij} representa la $j-esima$ componente del $i-esimo$ vector propio de M. Multiplicando U por M a izquierda se tiene:

$$\begin{aligned} MU &= \begin{pmatrix} m_{11} & m_{21} & \cdots & m_{n1} \\ m_{12} & m_{22} & \cdots & m_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ m_{1n} & m_{2n} & \cdots & m_{nn} \end{pmatrix}_{n\times n} \begin{pmatrix} u_{11} & u_{21} & \cdots & u_{n1} \\ u_{12} & u_{22} & \cdots & u_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1n} & u_{2n} & \cdots & u_{nn} \end{pmatrix}_{n\times n} \\ &= \begin{pmatrix} M\begin{pmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1n} \end{pmatrix} & M\begin{pmatrix} u_{21} \\ u_{22} \\ \vdots \\ u_{2n} \end{pmatrix} & \cdots & M\begin{pmatrix} u_{n1} \\ u_{n2} \\ \vdots \\ u_{nn} \end{pmatrix} \end{pmatrix}_{n\times n} \\ &= \begin{pmatrix} \vdots & \vdots & & \vdots \\ \mu_1 u_1 & \mu_2 u_2 & \cdots & \mu_n u_n \\ \vdots & \vdots & & \vdots \end{pmatrix} \\ &= \begin{pmatrix} \vdots & \vdots & & \vdots \\ u_1 & u_2 & \cdots & u_n \\ \vdots & \vdots & & \vdots \end{pmatrix} \begin{pmatrix} \mu_1 & & & \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_n \end{pmatrix} \\ &= UD \end{aligned}$$

donde D es una matriz diagonal cuyos elementos son los autovalores de la matriz M. Así se establece:

$$MU = UD \tag{1.2.2}$$

Dado que U es una matriz ortogonal, es decir, las filas forman una base ortonormal lo mismo que sus columnas, tenemos que las entradas de U^TU vienen dadas como:

$$\begin{cases} 1, & \text{si } i = j; \\ 0, & \text{si } i \neq j. \end{cases} \qquad i, j = 1, 2, \ldots, n.$$

Teniendose así que $U^TU = I$; además si aplicamos traspuesta en ambos lados obtenemos que $UU^T = I$, implicando esto que $U^T = U^{-1}$. Con esto aseguramos que la matriz

U es invertible y su inversa es su traspuesta.
Multiplicando en (1.2.2) por U^{-1} a derecha

$$MUU^{-1} = UDU^{-1} \Rightarrow M = UDU^{-1}$$

quedando,

$$M = UDU^T \tag{1.2.3}$$

lo cual se puede escribir como:

$$M = \sum_{k=1}^{n} \mu_k u_k u_k^T \tag{1.2.4}$$

De la forma dada en (1.2.4) podemos entender la aplicación de M a un vector $x \in \mathbb{R}$ de la siguiente manera:

Se resuelve el vector de entrada x a lo largo de cada autovector, la componente del vector de entrada a lo largo del k-ésimo autovector viene dada por $u_k^T x$. Así,

$$Mx = (\sum_{k=1}^{n} \mu_k u_k u_k^T)x = (\sum_{k=1}^{n} \mu_k u_k u_k^T x)$$

Notese que el producto $u_k^T x$ es la proyección del vector x a lo largo del vector u_k^T; dicho valor se multiplica por μ_k y el u_k presente en la expresión indica la dirección la cual se mantiene sobre u_k. En la siguiente figura se ilustra la situación para dos dimensiones.

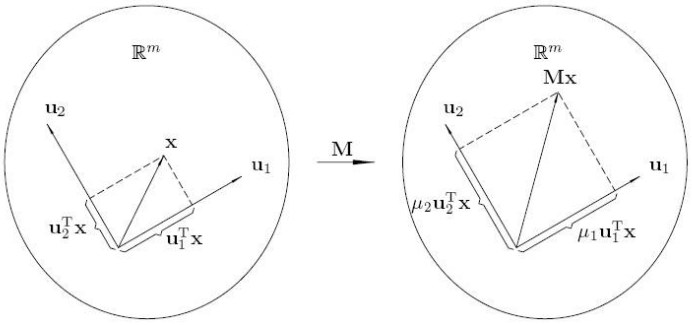

FIGURA 1.2.1: La aplicación $M : \mathbb{R}^m \to \mathbb{R}^m$

§1.2.2. Descomposición del valor singular de una matriz real rectangular

Consideremos ahora el caso más general de una matriz rectangular. Sea $A \in \mathbb{R}^{m\times n}$ (matriz de orden m por n), la cual envía vectores de $\mathbb{R}^n$ en vectores de $\mathbb{R}^m$ (espacio al que pertenecen las filas al espacio donde se encuentran las columnas).

Tenemos que tanto AA^T como A^TA son matrices cuadradas y fácilmente se verifica que son simétricas, donde $AA^T \in \mathbb{R}^{m\times m}$ y $A^TA \in \mathbb{R}^{n\times n}$; así, se puede hablar de autovalores y autovectores asociados a estas matrices. Con los autovectores obtenemos bases para los espacios respectivos.
Dado que $x^T(A^TA)x = (Ax)^T(Ax)$ para cualquier $x \in \mathbb{R}^n$, la operación denota un producto interno y por tanto el valor resultante es no negativo. Esto quiere decir que A^TA es una matriz definida positiva. De la misma forma:

$$x^T(AA^T)x = x^T(AA^T)^Tx = x^TA^TAx = (Ax)^T(Ax),$$

(la primera igualdad es por la simetría de AA^T)
quedando similar al caso anterior y obteniendose así que AA^T también es simétrica definida positiva.

Considerando v un autovector de A^TA y λ el autovalor correspondiente se tiene,

$$(A^TA)v = \lambda v, \tag{1.2.5}$$

entonces $v^T(A^TA)v = \lambda v^Tv \geqslant 0$ (A^TA es definida positiva), y como $v^Tv \geqslant 0$ no queda otra opción para λ que ser no negativo; con esto hemos probado que los autovalores de las matrices A^TA y AA^T son no negativos.

Como se dijo anteriormente la matriz A^TA es una matriz cuadrada de orden n, como consecuencia tiene n vectores ortonormales propios; denotemos cada uno de estos como v_i con $i = 1, 2, \ldots, n$ y los correspondientes autovalores λ_i, asumiendo además que estos han sido ordenados de manera que $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n \geqslant 0$.

Con un argumento similar, podemos considerar los vectores propios ortonormales de $AA^T \in \mathbb{R}^{m\times m}$ como los u_i con $i = 1, 2, \ldots, m$ y asumiendo los autovalores $\mu_1 \geqslant \mu_2 \geqslant \ldots \geqslant \mu_m \geqslant 0$.

Consideremos ahora el autovalor v_1 de A^TA y supongamos que λ_1 es no nulo.

Tenemos que $A^TAv_1 = \lambda_1 v_1 \neq 0 \quad (\lambda_1 \neq 0, v_1 \neq 0)$; pero si $A^TAv_1 \neq 0$, entonces Av_1 ha de ser no nulo.

Ahora mostraremos que el vector Av_1 es en realidad un autovector de la matriz $AA^T \in \mathbb{R}^{m\times m}$. (Notar que es la otra matriz en estudio).

$$(AA^T)Av_1 = A(A^TA)v_1 = A\lambda_1 v_1 = \lambda_1(Av_1), \tag{1.2.6}$$

donde λ_1 es el autovalor correspondiente y es además autovalor de A^TA.

Al normalizar el autovector Av_1 obtenemos $\frac{Av_1}{\|Av_1\|}$ (ya que se aseguró que $Av_1 \neq 0$).

Entonces $\frac{Av_1}{\|Av_1\|}$ es un vector propio normalizado de AA^T y por tanto ha de ser uno de los u_i mencionados arriba.

Haciendo las mismas consideraciones con cada uno de los λ_i restantes se llega a la conclusión que cada λ_i es autovalor de AA^T.

Con un procedimiento similar al realizado con λ_1 y v_1 pero ahora tomando en cuenta un autovector u_i de AA^T, se muestra que $\frac{A^Tu_i}{\|A^Tu_i\|}$ es también un autovector normalizado de A^TA con μ_i el autovalor correspondiente.

Así, se garantiza que los autovalores no nulos de A^TA son también autovalores de AA^T y viceversa.

Si existen r de n autovalores no nulos, lo anterior nos dice que $\lambda_1 = \mu_1$,$\lambda_2 = \mu_2$,...,$\lambda_r = \mu_r$ siendo el resto de los autovalores iguales a cero.

Así, resulta que para todo $k = 1, 2, \ldots, r$

$$u_k = \frac{Av_k}{\|Av_k\|} \qquad v_k = \frac{A^Tu_k}{\|A^Tu_k\|} \tag{1.2.7}$$

Esto pasa automaticamente si los autovalores no-nulos de A^TA y AA^T son no degenerados. Si hay degeneraciones entre los autovalores, es posible tomar la combinación

lineal apropiada en el autoespacio degenerado tal que (1.2.7) se satisface.

Ya que,

$$\begin{aligned} \|Av_k\|^2 &= (Av_k)^T(Av_k) \quad k = 1, \ldots, r && (1.2.8)\\ &= v_k^T A^T A v_k && (1.2.9)\\ &= v_k^T \lambda_k v_k \qquad (v_k \text{ es autovector de } A^T A) && (1.2.10)\\ &= \lambda_k v_k^T v_k && (1.2.11)\\ &= \lambda_k \qquad (\text{los } v_i \text{ forman una base ortonormal}) && (1.2.12) \end{aligned}$$

se tiene que de la primera expresión en (1.2.7) queda:

$$u_k = \frac{Av_k}{\sqrt{\lambda_k}} \quad \Rightarrow \quad Av_k = \sqrt{\lambda_k} u_k \tag{1.2.13}$$

de manera similar se obtiene que $\|A^T u_k\|^2 = \mu_k$. Sustituyendo en la segunda expresión de (1.2.7):

$$v_k = \frac{A^T u_k}{\sqrt{\mu_k}} \quad \Rightarrow \quad A^T u_k = \sqrt{\mu_k} v_k \tag{1.2.14}$$

Pero ya habíamos visto anteriormente que $\lambda_k = \mu_k, \quad con \quad k = 1, 2, \ldots, r$; así, $\sqrt{\lambda_k} = \sqrt{\mu_k}$ y denotando al valor común como σ_k, escribimos (1.2.13) y (1.2.14) como sigue:

$$\begin{aligned} Av_k &= \sigma_k u_k && (1.2.15)\\ A^T u_k &= \sigma_k v_k \qquad k = 1, 2, \ldots, r && (1.2.16) \end{aligned}$$

Tanto en (1.2.15) como en (1.2.16) se observa el efecto de la transformación lineal A y A^T respectivamente. La matriz A envía un v_k de la base al vector $\sigma_k u_k$, es decir, que A es una aplicación que va de $\mathbb{R}^n$ en $\mathbb{R}^m$. Por otra parte, A^T envía el vector u_k de $\mathbb{R}^m$ en el vector $\sigma_k v_k$ de $\mathbb{R}^n$.

Recordemos que estamos bajo el supuesto de r autovalores no nulos para cada una de las matrices AA^T y $A^T A$; así, cuando tomamos $k > r$, el autovalor λ_k de $A^T A$ es

nulo. El mismo está asociado a un autovector v_k, obtenemos así que $A^TAv_k = 0$. Si en la última expresión se multiplica por v_k^T a izquierda se obtiene:

$$v_k^T A^T A v_k = 0 \;\Rightarrow\; (Av_k)^T(Av_k) = 0 \tag{1.2.17}$$

$$\Rightarrow\; \|Av_k\| = 0 \tag{1.2.18}$$

$$\Rightarrow\; Av_k = \overrightarrow{0} \tag{1.2.19}$$

De forma totalmente análoga se verifica que $A^Tu_k = \overrightarrow{0}$. En resumen,

$$Av_k = \overrightarrow{0} \quad k = r+1, \ldots, n \tag{1.2.20}$$

$$A^Tu_k = \overrightarrow{0} \quad k = r+1, \ldots, m \tag{1.2.21}$$

Hemos así logrado con estas expresiones y con (1.2.15) y (1.2.16) obtener de manera explicita la acción que realiza la matriz A y la matriz A^T respectivamente sobre las bases $\{v_k\}_{k=1}^n$ y $\{u_k\}_{k=1}^m$.

Ahora bien, si se tiene cualquier operador lineal que haga la misma acción de A sobre los vectores de la base $\{v_k\}_{k=1}^n$, este operador ha de ser el mismo A. En efecto, solo basta ver que este nuevo operador realice la misma acción de la matriz A para todo vector en $\mathbb{R}^n$. Con los vectores básicos ya se satisface, y para cualquier otro vector w este es posible escribirlo como combinación lineal de la base $\{v_k\}_{k=1}^n$ y es claro que tanto A como el otro supuesto operador haran exactamente lo mismo, de donde se tiene el resultado.

Se quiere ahora obtener una expresión explícita para el operador A, por tanto, probemos que el operador $\sum_{k=1}^r \sigma_k u_k v_k^T$ realiza la misma acción de A.

Para $i = 1, \ldots, r$

$$(\sum_{k=1}^{r} \sigma_k u_k v_k^T)v_i = \sum_{k=1}^{r} \sigma_k u_k v_k^T v_i \quad \text{(Por linealidad)} \tag{1.2.22}$$

$$= \sigma_k u_i \tag{1.2.23}$$

La última igualdad se satisface ya que $v_k^T v_i = \begin{cases} 1, & \text{si } k = i; \\ 0, & \text{si } k \neq i. \end{cases}$

Para $i = r+1, \ldots, n$

$$(\sum_{k=1}^{r} \sigma_k u_k v_k^T) v_i = \sum_{k=1}^{r} \sigma_k u_k v_k^T v_i \qquad \text{(Por linealidad)} \tag{1.2.24}$$

$$= 0 \qquad \text{ya que } i \neq k \tag{1.2.25}$$

Por el resultado anterior se tiene que este operador es el mismo A. Así,

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^T \tag{1.2.26}$$

Aplicando traspuesta en ambos lados de (1.2.26) y por la linealidad tenemos:

$$A^T = \sum_{k=1}^{r} \sigma_k v_k u_k^T \tag{1.2.27}$$

DEFINICIÓN 1.2.1. Los vectores $\{v_k\}$ son denominados vectores singulares a derecha. Los vectores $\{u_k\}$ se les denomina vectores singulares a izquierda, y los escalares $\{\sigma_k\}$ son los valores singulares de la matriz.

De (1.2.26) se puede escribir $A = USV^T$, donde U es la matriz cuya k-ésima columna es u_k, V la matriz cuya k-ésima columna es v_k y $S_{m\times n}$ con entradas no nulas sólo en los primeros r-elementos de la diagonal con $s_{kk} = \sigma_k$.

§1.2.3. Interpretación de la descomposición del valor singular de una matriz

Recordemos que el problema inverso en estudio es de identificar la imagen f, dada la data d y la aplicación directa A. De este modo, nos gustaría saber que sentido toma Af cuando ya tenemos una expresión de A como la dada en (1.2.26).

De (1.2.26) tenemos:

$$Af = (\sum_{k=1}^{r} \sigma_k u_k v_k^T) f \tag{1.2.28}$$

$$= \sum_{k=1}^{r} \sigma_k u_k (v_k^T f) \tag{1.2.29}$$

Podemos describir el proceso que pasa f por A. $f \in \mathbb{R}^n$, esto significa que f puede ser escrito como combinación lineal de $\{v_k\}_{k=1}^r$. El producto $v_k^T f$ dado en (1.2.29) expresa

el valor o la componente de f a lo largo del vector v_k, esta última es multiplicada por el valor singular σ_k y finalmente por el vector u_k, que para efectos de la expresión (1.2.29) indica que la componenete de f sobre v_k ha sido enviada a la componente de u_k de Af.

El resultado se hace aún más significativo por el hecho que los conjuntos $\{v_k\}_{k=1}^r$, $\{u_k\}_{k=1}^r$ pueden ser extendidos de manera de obtener una base para los espacios respectivos y por que la base del espacio de llegada es distinto al espacio de salida (recordar que A es una matriz rectangular).

En el siguiente gráfico se ilustra la situación sobre dos dimensiones.

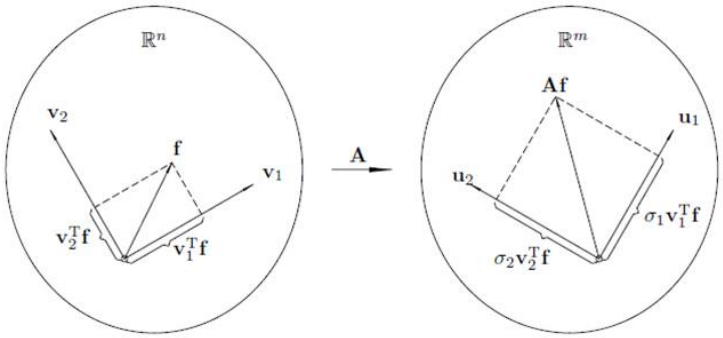

FIGURA 1.2.2: Efecto de una matriz rectangular $A \in \mathbb{R}^{m\times n}$ sobre un vector $f \in \mathbb{R}^n$

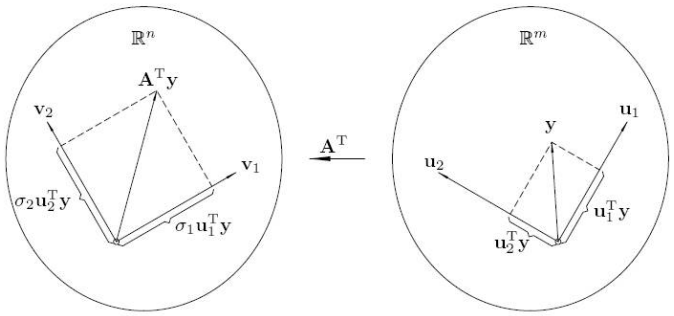

FIGURA 1.2.3: Efecto de una matriz rectangular $A^T \in \mathbb{R}^{n\times m}$ sobre un vector $y \in \mathbb{R}^m$

De manera análoga se estudia el efecto de A^T sobre un vector $y \in \mathbb{R}^m$ con el resultado obtenido en (1.2.27), pudiendose representar como en la figura (1.2.3).

§1.2.4. Geometría de una transformación lineal

Las expresiones (1.2.15) y (1.2.21) indican que $Imag(A) = span\{u_1, u_2, \ldots, u_r\}$ y $ker(A^T) = span\{u_{r+1}, u_{r+2}, \ldots, u_m\}$ respectivamente. Dado que $\{u_1, u_2, \ldots, u_m\}$ resulta ser una base para $\mathbb{R}^m$, este espacio puede escribirse como:

$$\mathbb{R}^m = Imag(A) \oplus ker(A^T). \tag{1.2.30}$$

Y por el mismo razonamiento:

$$\mathbb{R}^n = Imag(A^T) \oplus ker(A). \tag{1.2.31}$$

Estudiemos ahora como es que actúa el operador A sobre f para este ser enviado al espacio de datos. $f \in \mathbb{R}^n$ y tenemos como una base para $\mathbb{R}^n$ el conjunto de los vectores singulares a derecha $\{v_k\}_{k=1}^n$; podemos por tanto escribir:

$$f = \sum_{k=1}^{n} f_k v_k = \sum_{k=1}^{n} \langle f, v_k^T \rangle v_k, \qquad f_k = \langle f, v_k^T \rangle \tag{1.2.32}$$

esto no es más que la suma vectorial de las componentes de f a lo largo de las direcciones básicas, (1.2.29) se escribe como:

$$Af = \sum_{k=1}^{r} \sigma_k f_k u_k \tag{1.2.33}$$

La expresión anterior nos indica que en la data se encuentra información "codificada", por decirlo de alguna manera, de la imagen f; en concreto las proyecciones de esta a lo largo de los vectores v_k. Es importante notar de la expresión (1.2.33) la influencia que ejerce el valor σ_k para indicar que tan presente se encuentra la k-ésima componente de la imagen en la data.

Otro aspecto de relevancia es el hecho de venir trabajando con el valor r, el cual hemos dicho indica el rango de la matriz A; esto sumado a que $r < n$ señala que sólo las primeras r componentes de f tienen presencia en la data. Esto indica que la información suministrada sólo por la data es débil, en el sentido que, si se tienen dos imagenes distintas f y f' las cuales son iguales en las primeras r componentes, difiriendo en las restantes $n-r$ componentes, entonces para la misma data hallamos dos imagenes que satisfacen con la misma y son distintas, recordemos que queremos encontrar una

f que es única, y trabajando solo con la data podríamos tener varias soluciones para la misma data. Se hace necesario agregar técnicas que nos permitan determinar las componentes invisibles a la data.

§1.2.5. La descomposición del valor singular en problemas de modelos de ajuste

En problemas de modelos de ajuste, tales como el de fijar una recta $y = f_0 + f_1 x$ en una colección de m datos (puntos) $\{(x_k, y_k)\}_{k=1}^m$, aquí la dimensión del espacio imagen es bastante pequeña. El problema $\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \end{pmatrix}$, es tal que el espacio imagen (no confundir con la imagen de la matriz) es de dimensión 2 $(n = 2)$, mientras que el de la dimensión del espacio de datos es m. Por la naturaleza del problema, el problema arroja respuestas bien definidas cuando el rango de la matriz es igual a n (en este caso 2); así la imagen de A es un subespacio dos dimensional de $\mathbb{R}^m$. Los datos que están en la imagen de A son aquellos que para el modelo de ajuste los puntos quedan exactamente sobre la recta que este modela. Si se tienen más datos, los puntos no estarán todos sobre la misma linea recta y se tendrá que $d \notin img(A)$.

Cuando se considera la aproximación de los mínimos cuadrados, esta busca tomar $\widehat{f}$ (parámetro) tal que Af se aproxime tanto como sea posible a $\mathbf{d}$, es decir:

$$\widehat{f} = \arg\min \|d - Af\|^2 \tag{1.2.34}$$

Apliquemos lo que sabemos de la descomposición del valor singular de una matriz A. En (1.2.26) vimos que la matriz A puede ser escrita como:

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^T, \tag{1.2.35}$$

donde σ_k es la raiz cuadrada del k-ésimo autovalor de AA^T, ordenados estos de manera decreciente; u_k es el k-ésimo autovalor respectivo de AA^T y v_k es el traspuesto del k-ésimo autovector de $A^T A$.

Ahora bien, en nuestro caso actual, consideraremos la data de tamaño m y como (1.2.33) nos indica que la data puede ser escrita como una combinación lineal de la base

$\{u_k\}_{k=1}^m$. Así, la data se escribe como la suma vectorial de las componentes a lo largo de cada vector básico, quedando:

$$d = \sum_{k=1}^{m} u_k(u_k^T d) \tag{1.2.36}$$

El problema de minimizar $\|d - Af\|^2$ para un f dado puede ser estudiado como sigue:

$$\begin{aligned} \|d - Af\|^2 &= \|\sum_{k=1}^{m} u_k(u_k^T d) - \sum_{k=1}^{r} \sigma_k u_k(v_k^T f)\|^2 \quad \text{(por (1.2.26) y (1.2.36))} && (1.2.37)\\ &= \|\sum_{k=1}^{r}[u_k(u_k^T d) - \sigma_k u_k(v_k^T f)] + \sum_{k=r+1}^{m} u_k(u_k^T d)\|^2 && (1.2.38)\\ &= \|\sum_{k=1}^{r} u_k[(u_k^T d) - \sigma_k(v_k^T f)] + \sum_{k=r+1}^{m} u_k(u_k^T d)\|^2 && (1.2.39) \end{aligned}$$

Como dentro de las barras de norma la suma involucrada es de vectores ortonormales ($u_i \perp u_j$ para $i \neq j$, el teorema general de Pitágoras nos garantiza,

$$\begin{aligned} &\|\sum_{k=1}^{r} u_k[(u_k^T d) - \sigma_k(v_k^T f)] + \sum_{k=r+1}^{m} u_k(u_k^T d)\|^2 \\ &\qquad = \|\sum_{k=1}^{r} u_k[(u_k^T d) - \sigma_k(v_k^T f)]\|^2 + \|\sum_{k=r+1}^{m} u_k(u_k^T d)\|^2 \\ &\qquad\qquad = \sum_{k=1}^{r} |(u_k^T d) - \sigma_k(v_k^T f)|^2 + \sum_{k=r+1}^{m} |u_k^T d|^2 \quad (1.2.40) \end{aligned}$$

La última igualdad se obtiene por la ortonormalidad de los u_k, $k = 1, 2, \ldots, m$.

Consideremos $\widehat{f}$ el "vector imagen"que minimiza la expresión $\|Af - d\|^2$.

El segundo término de la derecha de (1.2.40) no depende del "vector imagen"tomado, el término restante si; y este es mínimo cuando $\widehat{f}$ es tomado de manera que:

$$v_k^T \widehat{f} = \frac{u_k^T d}{\sigma_k} \quad \text{para } k = 1, 2, \ldots, r \tag{1.2.41}$$

Que la expresión determine o no a $\widehat{f}$ depende de si $r = n$ o $r < n$. Para el modelo de ajuste y considerando $r = n$ la única solución al problema es:

$$\widehat{f} = \sum_{k=1}^{n}(v_k^T \widehat{f})v_k \qquad \text{(por (1.2.32))} \tag{1.2.42}$$

$$= \sum_{k=1}^{n}\left(\frac{u_k^T d}{\sigma_k}\right)v_k \qquad \text{(Considerando (1.2.41))} \tag{1.2.43}$$

$$= \sum_{k=1}^{n} v_k\left(\frac{u_k^T d}{\sigma_k}\right) \tag{1.2.44}$$

$$= \left(\sum_{k=1}^{n}\frac{1}{\sigma_k}v_k u_k^T\right)d \tag{1.2.45}$$

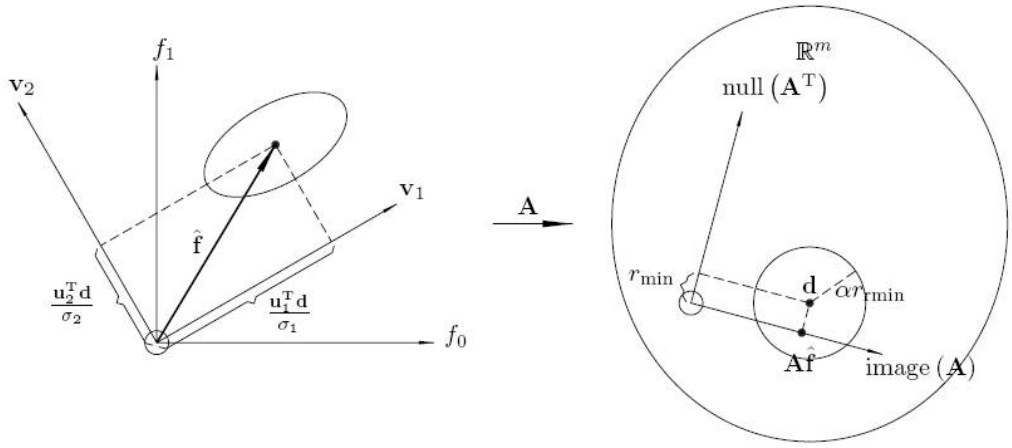

FIGURA 1.2.4: Geometría del problema de modelo de ajuste

Este proceso para el problema de ajuste se ilustra en la figura (1.2.4) de forma esquemática. El espacio imagen que es de dos dimensiones se representa de forma completa, mientras que $imag(A)$ y $ker(A^T)$ son representados por dos lineas perpendiculares. La data d se muestra ligeramente fuera de $imag(A)$ debido al ruido. Los parametros $\widehat{f}$ son tomados tal que $A\widehat{f}$ sea cercano como sea posible a la data d en el espacio de datos. Las componentes de $\widehat{f}$ a lo largo de los vectores v_k vienen dados por $u_k^T d/\sigma_k$ como ya se había indicado en (1.2.45). Los vectores singulares v_1, v_2, al formar una base del espacio de imagenes, forman un ángulo recto entre sí y estos a su vez hacen ángulo con los ejes f_0, f_1, que representan la intersección y derivada de la linea recta que se estima. La elipse en el espacio imagen indica la incertidumbre en los valores del parámetro fijado los cuales se detallan más adelante.

Relación con la inversa Moore-Penrose

Para el problema de minimizar $\|d - Af\|^2$, podemos escribir

$$C = \|d - Af\|^2 \tag{1.2.46}$$

y como $d - Af$ es un vector, por definición su norma al cuadrado no es más que la sumatoria de sus componentes al cuadrado. Por una parte, $\sum_{l=1}^{n} a_{kl} f_l$ es la k-ésima componente de Af. Así, podemos escribir

$$C = \sum_{k=1}^{m} \left(d_k - \sum_{l=1}^{n} a_{kl} f_l \right)^2 \tag{1.2.47}$$

Como se trata de minimizar la expresión, se considera el mínimo para cada componente,

$$\frac{\partial C}{\partial f_i} = \sum_{k=1}^{m} 2 \left(d_k - \sum_{l=1}^{n} a_{kl} f_l \right) (-a_{ki}) = 0, \quad \text{i=1,2,...,n} \tag{1.2.48}$$

$$\Rightarrow \quad 2 \sum_{k=1}^{m} \left(-a_{ki} d_k + \sum_{l=1}^{n} a_{kl} a_{ki} f_l \right) = 0 \quad \text{i=1,2,...,n} \tag{1.2.49}$$

$$\Rightarrow \quad - \sum_{k=1}^{m} a_{ki} d_k + \sum_{l=1}^{n} \left(\sum_{k=1}^{m} a_{ki} a_{kl} \right) f_l = 0 \tag{1.2.50}$$

$$\Rightarrow \quad \sum_{l=1}^{n} \left(\sum_{k=1}^{m} a_{ki} a_{kl} \right) f_l = \sum_{k=1}^{m} -a_{ki} d_k \tag{1.2.51}$$

$$\Rightarrow \quad (A^T A) f = A^T d \tag{1.2.52}$$

La representación dada en (1.2.52) es la de las ecuaciones normales del problema de los mínimos cuadrados.

Es claro que se tendrá solución única cuando la matriz $A^T A$ sea invertible; cuando eso ocurra el $\widehat{f}$ que minimiza C estará dado como:

$$\widehat{f} = (A^T A)^{-1} A^T d \tag{1.2.53}$$

DEFINICIÓN 1.2.2. (Inversa Moore-Penrose)
Sea A una matriz de orden $m \times n$. La matriz $(A^TA)^{-1}A^T$ es la **inversa Moore-Penrose de A** [6]

Ahora veamos como está relacionada la inversa Moore-Penrose del operador A con la descomposición del valor singular del mismo. Por una parte,

$$AA^T = \left(\sum_{k=1}^{r} \sigma_k v_k u_k^T\right)\left(\sum_{l=1}^{r} \sigma_l u_l v_l^T\right) \quad \text{(por (1.2.26),(1.2.27))} \tag{1.2.54}$$

$$= \sum_{k=1}^{r}\sum_{l=1}^{r} \sigma_k\sigma_l v_k (u_k^T u_l) v_l^T \tag{1.2.55}$$

$$= \sum_{k=1}^{r} \sigma_k^2 v_k v_k^T \tag{1.2.56}$$

La última igualdad se justifica por el hecho que:

$$u_k^T u_l = \begin{cases} 1, & \text{si } k = l; \\ 0, & \text{si } k \neq l; \end{cases} \tag{1.2.57}$$

Recuerde que A^TA es una matriz cuadrada y que la expresión dada en (1.2.56) es la misma expresión obtenida en (1.2.4), indicandonos que $\{\sigma_k^2\}_{k=1}^r$ son los autovalores no nulos de A^TA y como $A^TA \in \mathbb{M}_{n\times n}(\mathbb{R})$, esta es invertible si y sólo si $r = n$. En este caso,

$$(A^TA)^{-1} = \sum_{k=1}^{r} \frac{1}{\sigma_k^2} v_k v_k^T \tag{1.2.58}$$

Así, la inversa Moore-Penrose se escribe como,

$$(A^TA)^{-1}A^T = \left(\sum_{k=1}^{r} \frac{1}{\sigma_k^2} v_k v_k^T\right)\left(\sum_{l=1}^{r} \sigma_l v_l u_l^T\right) \quad \text{(de (1.2.58) y (1.2.27))} \tag{1.2.59}$$

$$= \sum_{k=1}^{r}\sum_{l=1}^{l} \frac{1}{\sigma_k^2}\sigma_l v_k (v_k^T v_l) u_l^T \tag{1.2.60}$$

$$= \sum_{k=1}^{r} \frac{1}{\sigma_k} v_k u_k^T \tag{1.2.61}$$

De (1.2.61) tenemos que el resultado para $\widehat{f}$ es el mismo que cuando se trabaja sólo

con la S.V.D. (1.2.45) o cuando se trabaja con la inversa Moore-Penrose. (Sustituyendo (1.2.61) en (1.2.53)).

Efectos del ruido sobre estimados en modelo paramétrico

Ahora interesa tomar en cuenta que tan bueno es el estimador $\widehat{f}$ que se tiene; no hay que olvidar que $\|d - A\widehat{f}\|^2$ nos expresa numericamente que tan deseable es $\widehat{f}$. El mejor estimador $\widehat{f}$ es el que minimiza la norma anterior. Es decir,

$$\|d - A\widehat{f}\|^2 = \min_{f \in \mathbb{R}^n} \|d - A\widehat{f}\|^2 = r_{min}^2 \tag{1.2.62}$$

donde es claro que r_{min} nos expresa distancia entre la $imag A$y la data $\mathbf{d}$.

Con la intención de revisar que tan confiable es un estimador $\widehat{f}$ de f, consideremos

$$F = \{f : \|d - Af\|^2 \leqslant \alpha r_{min}^2\} \tag{1.2.63}$$

como el conjunto factible, formado por los valores f-probables que minimizan (1,2,63); donde α regula el nivel de confianza requerido.

Estudiemos el lugar geométrico que representa el conjunto F como subconjunto del espacio de las imágenes.

$$\|d - Af\|^2 = \sum_{k=1}^{r} |(u_k^T d) - \sigma_k(v_k^T f)|^2 + \sum_{k=r+1}^{m} |u_k^T d|^2 \tag{1.2.64}$$

pero el segundo término de la expresión anterior es la norma al cuadrado de la data a lo largo del kernel de A^T, o equivalentemente indica la distancia al cuadrado de la data $\mathbf{d}$ a la $imag(A)$, a saber, $A\widehat{f}$. Por tanto,

$$\begin{aligned} \|d - Af\|^2 &= \sum_{k=1}^{r} |(u_k^T d) - \sigma_k(v_k^T f)|^2 + r_{min}^2 && (1.2.65) \\ &= \sum_{k=1}^{r} |\sigma_k(v_k^T \widehat{f}) - \sigma_k(v_k^T f)|^2 + r_{min}^2 \quad \text{(por (1.2.41))} && (1.2.66) \\ &= \sum_{k=1}^{r} \sigma_k^2 |v_k^T(\widehat{f} - f)|^2 + r_{min}^2 && (1.2.67) \end{aligned}$$

Ahora bien, nosotros consideramos los $f \in F$, es decir, que se satisfaga

$$\sum_{k=1}^{r} \sigma_k^2 |v_k^T(\widehat{f} - f)|^2 + r_{min}^2 \leqslant \alpha r_{min}^2 \tag{1.2.68}$$

$$\Rightarrow \sum_{k=1}^{r} \sigma_k^2 |v_k^T(\widehat{f} - f)|^2 \leqslant r_{min}^2(\alpha - 1) \tag{1.2.69}$$

La frontera del conjunto de f que satisfacen la expresión anterior en una elipse centrada en $\widehat{f}$ con ejes principales en las direcciones de los v_k. La longitud de k-ésimo semi-eje es: $\frac{r_{min}\sqrt{\alpha-1}}{\sigma_k}$

§1.2.6. Los efectos del ruido y pequeños valores singulares

Es importante observar la influencia que ejercen los valores singulares pequeños, en las direcciones de sus respectivos vectores propios durante el proceso de reconstrucción de la imagen.

Recordemos que la data puede ser expresada en términos de la imagen verdadera y el ruido, por la expresión

$$d = Af + n \tag{1.2.70}$$

de la descomposición del valor singular del operador A se obtiene

$$d = (\sum_{k=1}^{n} \sigma_k u_k v_k^T) f + n \tag{1.2.71}$$

Donde se ha supuesto el rango igual a n, tamaño del espacio imagen (a su vez menor que m). La aplicación directa es $1-1$; así, que el problema tiene solución única en el sentido de los mínimos cuadrados y que se indicó que venía dada por:

$$\widehat{f} \;=\; \left(\sum_{k=1}^{n} \frac{1}{\sigma_k} v_k u_k^T \right) d \tag{1.2.72}$$

Si en esta última expresión, sustituimos la expresión (1.2.71), obtenemos:

$$\widehat{f} = \sum_{k=1}^{n} \left(\frac{1}{\sigma_k} u_k^T \mathbf{d} \right) v_k \tag{1.2.73}$$

$$= \sum_{k=1}^{n} \left(\frac{1}{\sigma_k} u_k^T \left[\sum_{k=1}^{n} \sigma_k u_k v_k^T) f + n \right] \right) v_k \tag{1.2.74}$$

$$= \sum_{k=1}^{n} \left(\frac{1}{\sigma_k} u_k^T \sigma_k u_k v_k^T f + \frac{u_k^T \mathbf{n}}{\sigma_k} \right) v_k \tag{1.2.75}$$

$$= \sum_{k=1}^{n} \left(v_k^T f v_k + \frac{u_k^T \mathbf{n}}{\sigma_k} v_k \right) \quad (\{u_k\} \text{ es base ortonormal}) \tag{1.2.76}$$

$$= f + \sum_{k=1}^{n} \frac{1}{\sigma_k} (u_k^T \mathbf{n}) v_k \quad (\{v_k\} \text{ es base ortonormal}) \tag{1.2.77}$$

En esta última expresión, se observa claramente que, la reconstrucción consiste de la imagen real más un término indicando la presencia del ruido. Esto nos está indicando que el ruido presente en la k-ésima componente de la reconstrucción que se hace, depende del ruido presente en la k-ésima componente de la data, dividido por el valor singular correspondiente. De aquí, que el ruido es amplificado a lo largo de las componentes donde los valores singulares son pequeños; obteniendose que la componente de f a lo largo de esta dirección sea opacada. En otras palabras, podríamos decir que la data contiene poca información de la imagen a lo largo de esas componentes. Si se intenta amplificar la señal -por decirse de alguna manera- que aporta la data, inevitablemente se estaría amplificando el ruido presente. De lo que se puede decir, que el método de los mínimos cuadrados nos da malas reconstrucciones cuando se tienen valores singulares muy pequeños. Una alternativa en esta situación, es considerar estos valores singulares pequeños iguales a cero, y trabajar las componentes de f respectivas, como parámetros libres o independientes de la data, buscando reconstruir estos por otros métodos, como se ilustrará más adelante.

§1.3. Métodos de Regularización para problemas inversos lineales

Debido a que no se cuenta con toda la información sobre la f, debemos considerar información adicional que nos permita tener criterios de selección dentro de un conjunto factible de reconstrucciones.

Una forma puede ser introduciendo una nueva función que busque una aproximación a una solución f^∞ que llamaremos solución por defecto, la cual en la práctica se obtiene a menudo de la información histórica del problema. Así, se puede considerar una nueva función

$$\Omega(f) = \|f - f^\infty\|^2 \tag{1.3.1}$$

con la cual se desea tener una aproximación a la f. En contextos más generales, se consideran operadores actuando sobre la diferencia $f - f^\infty$; considerando un operador L se puede escribir

$$\begin{aligned} \Omega(f) &= \|L(f - f^\infty)\|^2 && (1.3.2)\\ &= [L(f - f^\infty)]^T L(f - f^\infty) && (1.3.3)\\ &= (f - f^\infty)^T L^T L(f - f^\infty) && (1.3.4) \end{aligned}$$

Lo anterior es de utilidad, cuando se trabaja con el siguiente método de regularización.

§1.3.1. Regularización de Tickhonov

Se realiza considerando una suma ponderada de las funciones $C(f) = \|d - Af\|^2$ y $\Omega(f) = \|L(f - f^\infty)\|^2$. Como ya se ha indicado, la primera de las funciones nos da el aporte que viene de la data, sin evitar la componente de ruido presente; la segunda por su parte, proviene de información experimental de caracter histórico.

Así, considerando además un factor de ponderación λ^2, deseamos encontrar $\widehat{f_\lambda}$ que minimiza la suma ponderada, es decir:

$$\widehat{f_\lambda} = \arg\min\{\lambda^2\|L(f - f^\infty)\|^2 + \|d - Af\|^2\} \tag{1.3.5}$$

DEFINICIÓN 1.3.1. El factor de ponderación λ^2 es definido como parámetro de regularización.

Es clara la dependencia de la solución respecto al factor λ^2, tanto si este se hace tender a infinito, como si se hace tender a cero. Para lo primero, la información proveniente de la data y por ende del ruido presente, se hace insignificante, tomando valor nuestra solución por defecto, resultando ser esta la $\widehat{f_\lambda}$ buscada. En el caso que $\lambda \to 0$, la solución por defecto influye poco, obteniendo mucha importancia la información que

se extrae de la data. El objetivo es por tanto, controlar el λ adecuado, de forma de obtener un resultado óptimo.

La expresión,

$$\lambda^2\|L(f-f^\infty)\|^2+\|d-Af\|^2 \tag{1.3.6}$$

puede escribirse como:

$$\lambda^2(f-f^\infty)^TL^TL(f-f^\infty)+(d-Af)^T(d-Af) \tag{1.3.7}$$

y querer obtener el argumento mínimo es equivalente a que:

$$\frac{\partial}{\partial f_k}(\lambda^2(f-f^\infty)^TL^TL(f-f^\infty)+(d-Af)^T(d-Af))=0, \qquad k=1,2,\ldots,n \tag{1.3.8}$$

El problema se reduce a resolver el sistema:

$$\begin{aligned} 2\lambda^2L^TL(f-f^\infty)-2A^T(d-Af)=0 &\Rightarrow 2\lambda^2L^TLf-2\lambda^2L^TLf^\infty-2A^Td+2A^TAf=0 \\ &\Rightarrow (2\lambda^2L^TL+2A^TA)f=2\lambda^2L^TLf^\infty+2A^Td \quad (1.3.9)\\ &\Rightarrow (\lambda^2L^TL+A^TA)f=\lambda^2L^TLf^\infty+A^Td \quad (1.3.10)\end{aligned}$$

Obtendremos solución única en el sistema en la medida que $\lambda^2L^TL+A^TA$ sea no singular; donde además ya hemos probado que una matriz de esa forma es simétrica definida positiva.

§1.3.2. Descomposición del valor singular truncado

El método se basa en que al tenerse ya la SVD del operador A y estar dado como:

$$A=\sum_{l=1}^{r}\sigma_lu_lv_l^T \tag{1.3.11}$$

se observe los valores singulares más grandes de A para los cuales las componentes de la reconstrucción de f a lo largo de los respectivos vectores singulares estén bien determinados por la data. Se toma un entero $k\leqslant n$, considerando los valores singulares σ_k con $k>n$ ser de valores muy pequeños y el vector solución $\widehat{f}$ es tomado tal que

$$v_l^T\widehat{f}=\frac{u_l^Td}{\sigma_l} \quad \text{para } l=1,2,\ldots,k. \tag{1.3.12}$$

Las componentes a lo largo de los restantes direcciones $\{v_l\}$, l=k+1,...,n son tomadas de forma que el vector solución total $\widehat{f}$ satisfaga algún criterio de optimalidad, tal como la minimización de una función como la $\Omega(f)$ indicada antes.

Capítulo 2

Problemas Inversos en Estadística no Paramétrica

§2.1. Problemas Inversos en Estadística no Paramétrica

§2.1.1. Introducción

Tal como es presentado en [6], en pocas palabra podemos decir que resolver un problema inverso consiste en la reconstrucción de unos parámetros, que es lo que se ha considerado imagen, y que hemos estado denotando por f. La reconstrucción se realiza a partir de unas observaciones realizadas en las cuales hay presencia de ruido. Los datos que deseamos obtener son modificados por un operador que se aplica a f, enviando esta al espacio de datos.

Se presentan dos dificultades al momento de reconstruir la f: Considerando como modelo para el problema a la expresión $Af = d$, donde d es la data sin error; por una parte, la data que medimos no es exactamente Af, es decir, tenemos una versión ruidosa de Af y por otra parte que para muchos problemas de interés, la aplicación A no es invertible o cercana a no serlo, cuestión que merece atención en los problemas inversos. Nos encontramos así ante lo que hemos llamado problemas inversos mal puestos.

Nos referimos a problemas inversos en estadística cuando al menos una de las componentes del problema tiene un trato bajo un enfoque estocástico, donde usualmente es el ruido el que se trabaja en este enfoque. Entonces uno de los objetivos es estudiar métodos de regularización estadística que permitan una reconstrucción significativa a pesar del mal posicionamiento y el ruido.

El objetivo del paper en estudio es explicar algunos temas teóricos básicos relaciona-

dos a la estructura estadística de los problemas inversos.

Las siguientes son unas definiciones previas, relacionadas con los espacios en los cuales trabajaremos para desarrollar la teoría de aquí en adelante.

DEFINICIÓN 2.1.1 (Espacio de Hilbert)**.** Sea $(H, \langle\cdot\rangle)$ un espacio con producto interno y sea $\|\cdot\|$ su norma asociada. Si $(H, \|\cdot\|)$ es un espacio normado completo, entonces decimos que $(H, \langle\cdot\rangle)$ es un **espacio de Hilbert**. [8]

DEFINICIÓN 2.1.2. Sea E y F dos espacios normados cualesquiera. Consideremos un operador lineal (Transformación lineal continua entre dos espacios normados) $T : E \to F$, se dice que T es un **operador acotado** si T es acotado en la esfera unitaria $S = \{x \in E : \|x\| \leqslant 1\}$ de E, o sea, si existe "a" fijo tal que $\|Tx\| \leqslant a, \quad \forall x \in S$. Entre tales constantes "a" existe una que es mínima y recibe el nombre de **norma de** T denotada como $\|T\|$. [8]

La razón por la que se dan las definiciones anteriores es porque la data que dispondremos en los problemas son una versión ruidosa de Af, donde f es una función desconocida perteneciendo a un espacio con producto interno y A un operador lineal acotado entre dos espacios de Hilbert separables.

Que los espacios sean separables nos garantiza que dicho espacio tiene bases ortonormales numerables [8]; esto tendrá su importancia cuando describamos la descomposición del valor singular en este espacio.

La estructura estandar para problemas inversos corresponde al modelo con ruido determinístico (y aditivo), donde ξ es considerado pequeño. La aproximación que consideraremos acá de ahora en adelante es estadística: El ruido es considerado una variable aleatoria.

Se asumirá además el ruido blanco Gaussiano, lo cual significa que en el caso donde $G = \mathbb{R}^n$, el ruido es un vector de variables aleatorias Gaussianas independientes e igualmente distribuidas.

En general, ξ es el ruido blanco Gaussiano si para funciones cualesquiera $g_1, g_2 \in G$, las variables aleatorias $\langle \xi, g_j \rangle$ son $\mathcal{N}(0, \|g_j\|^2)$ con covarianza $(\langle \xi, g_1 \rangle, \langle \xi, g_2 \rangle) = \langle g_1, g_2 \rangle$, donde $\|\cdot\|$ y $\langle \cdot, \cdot \rangle$ representan respectivamente la norma y el producto escalar de H y G.

Una observación que vale la pena hacer es que, mientras el ruido determinístico se considera pequeño ($\|\xi\| \leqslant 1$) en el ámbito estadístico un proceso con ruido blanco este no pertenece al espacio H porque $\|\xi\| = \infty$.

El modelo estadístico de muestra discreta estandar para problemas lineales inversos es:

$$Y_i = Af(X_i) + \xi_i, \qquad i = 1, \ldots, n \tag{2.1.1}$$

donde $(X_1, Y_1), \ldots, (X_n, Y_n)$ son las observaciones (se puede asumir $X_i \in [0, 1]$), f es una función desconocida en $L^2(0, 1)$, A es un operador de $L^2(0, 1)$ en $L^2(0, 1)$ y ξ_i son variables aleatorias Gaussianas independientes igualmente distribuidas con media cero y varianza σ^2.

En estudios teóricos, se verá (2.1.1) a menudo escrito como un modelo de ruido blanco Gaussiano (es decir, proceso estocástico de ruido blanco).

$$Y = Af + \epsilon\xi \tag{2.1.2}$$

donde ξ es un error estocástico y $\epsilon > 0$ es el nivel de ruido. En el contexto de los problemas inversos (2.1.2) puede ser visto como una versión idealizada de (2.1.1), donde $\{y_i\}_{i=1}^n$ y $\{\xi_i\}_{i=1}^n$ son reemplazada por las funciones continuas y y ξ. Podemos ver este cambio como el proceso inverso de discretización finita; considerando además la equivalencia entre los modelos cuando $\epsilon \to 0$ y $n \to \infty$ respectivamente.

En algunos casos, como el considerado en ([6]), A no tiene el espacio nulo trivial, esto se describió cuando se consideró espacios euclideos. Allí vimos que las componentes de f que están en el espacio nulo no pueden ser resconstruidas sin más información y así la regularización del problema da más información sobre la función a reconstruir. Por simplicidad, se considerará la estimación de f cuando A es un operador inyectivo (pero no necesariamente estable).

De acuerdo al contexto en el que se trabaja se indican objetivos variados al hecho de obtener un estimado de f dada la data Y. En el campo de la estadística, es tratar con observaciones ruidosas para conseguir estimados de inversión y evaluar sus propiedades; en la teoría de problemas inversos, se trata de invertir el operador A y

en la matemática computacional buscar implementaciones numéricas para aplicaciones prácticas. De aquí en adelante nos fijaremos en la visión estadística y de la teoría de los problemas inversos, para describir algunas propiedades de este tipo de problemas.

Descomposición del valor singular y el modelo en el espacio de sucesiones

Nosotros discretizamos el modelo de ruido blanco al diagonalizar el operador directo A. Esto se hará usando la descomposición del valor singular.

Consideremos A^* el adjunto de A, el cual está definido como el traspuesto conjugado de A. Supongamos que A^*A es un operador compacto el cual tiene un sistema de autovectores ortogonales completo $\{\varphi_k\}$ con valores propios correspondientes ρ_k.

Recordemos la descomposición del valor singular presentada en el caso donde A se consideró un operador entre espacios euclideos. Acá de igual forma diremos que A admite una descomposición del valor singular cuando tengamos,

$$A^*Af = \sum_{k=1}^{\infty} b_k^2 \langle f, \varphi_k \rangle \varphi_k, \tag{2.1.3}$$

donde $\{b_k\}$ son los valores singulares, φ_k son los vectores singulares a derecha y $\langle f, \varphi_k \rangle$ representa las componentes de la imagen f con respecto a la base $\{\varphi_k\}$.

Se da ahora una definición que caracteriza el operador adjunto y que se usará para desarrollar algunos cálculos sencillos.

DEFINICIÓN 2.1.3. Sea H un espacio de Hilbert y sea A un operador lineal acotado. Un operador acotado B sobre H es llamado el adjunto de A si

$$\langle Ax, y \rangle = \langle x, By \rangle$$

para todo x, y en H. [8]

Definimos ahora la imagen normalizada $\{\psi_k\}$ de $\{\varphi_k\}$ como:

$$A\varphi_k = b_k \psi_k \tag{2.1.4}$$

de ahí podemos escribir,

$$\psi_k = b_k^{-1} A\varphi_k \tag{2.1.5}$$

y calculando su norma al cuadrado,

$$\begin{aligned} \|\psi_k\|^2 &= \langle b_k^{-1} A\varphi_k, b_k^{-1} A\varphi_k\rangle && (2.1.6)\\ &= b_k^{-2}\langle A\varphi_k, A\varphi_k\rangle && (2.1.7)\\ &= b_k^{-2}\langle A^*A\varphi_k, \varphi_k\rangle \qquad (\textit{Por definición de } A^*) && (2.1.8)\end{aligned}$$

Tenemos además que,

$$\begin{aligned} A^*A\varphi_k &= \sum_{j=1}^{\infty} b_j^2\langle\varphi_k, \varphi_j\rangle\varphi j \quad \textit{por } (2.1.3) && (2.1.9)\\ &= b_k^2\varphi_k \qquad \{\varphi_k\} \textit{ representa una base ortonormal.} && (2.1.10)\end{aligned}$$

Sustituyendo en (2.1.8),

$$\begin{aligned} \|\psi_k\|^2 &= b_k^{-2}\langle b_k^2\varphi_k, \varphi_k\rangle && (2.1.11)\\ &= b_k^{-2}b_k^2\langle\varphi_k, \varphi_k\rangle && (2.1.12)\\ &= \|\varphi_k\|^2 && (2.1.13)\\ &= 1, && (2.1.14)\end{aligned}$$

volviendo a (2.1.4),

$$\begin{aligned} A\varphi_k = b_k\psi_k &\Rightarrow A^*A\varphi_k = b_kA^*\psi_k && (2.1.15)\\ &\Rightarrow b_k^{-1}A^*A\varphi_k = A^*\psi_k && (2.1.16)\\ &\Rightarrow b_k^{-1}b_k^2\varphi_k = A^*\psi_k \qquad (\textit{por } (2.1.10)) && (2.1.17)\\ &\Rightarrow b_k\varphi_k = A^*\psi_k && (2.1.18)\end{aligned}$$

De (2.1.4) y (2.1.18) tenemos:

$$A\varphi_k = b_k\psi_k \qquad A^*\psi_k = b_k\varphi_k \tag{2.1.19}$$

Recordemos la expresión para nuestro modelo:

$$Y = Af + \epsilon\xi,$$

donde ξ es un error estocástico y $\epsilon > 0$ el nivel de ruido.

Aquí, podemos hablar de las componentes de Y a lo largo de los vectores básicos

$\{\psi_k\}$, hallemos su expresión explícita.

$$\begin{aligned} y_k &= \langle Y, \psi_k \rangle && (2.1.20)\\ &= \langle Af + \epsilon\xi, \psi_k \rangle && (2.1.21)\\ &= \langle Af, \psi_k \rangle + \epsilon\langle \xi, \psi_k \rangle && (2.1.22)\\ &= \langle Af, b_k^{-1} A\varphi_k \rangle + \epsilon\langle \xi, \psi_k \rangle \qquad (por\ (2.1.4)) && (2.1.23)\\ &= b_k^{-1}\langle Af, A\varphi_k \rangle + \epsilon\xi_k \qquad (con\ \xi_k = \langle \xi, \psi_k \rangle) && (2.1.24)\\ &= b_k^{-1}\langle A^*Af, \varphi_k \rangle + \epsilon\xi_k \qquad (\text{por definición } (2.1.3)) && (2.1.25)\\ &= b_k^{-1}\langle \sum_{k=1}^{\infty} b_k^2\theta_k\varphi_k, \varphi_k \rangle + \epsilon\xi_k \quad (\text{con } \theta_k = \langle f, \varphi_k \rangle) && (2.1.26)\\ &= b_k\theta_k + \epsilon\xi_k && (2.1.27) \end{aligned}$$

Ya que ξ es ruido blanco, tenemos que $\{\xi_k\}$ es una sucesión de variables independientes igualmente distribuidas con distribución $\mathcal{N}(0,1)$ (ver comentario antes de 2.1.1, con $\|\psi\|^2 = 1$). Así, tenemos un modelo de observaciones de sucesiones el cual es un equivalente discreto que se deriva de (2.1.2). El problema de estimar f dado Y, se transforma ahora en estimar la sucesión $\theta = \{\theta_k\}$ dada la sucesión discreta $y = \{y_k\}$.

La intención ahora es estimar $\{\theta_k\}$, el siguiente modelo es equivalente a (2.1.27) y resulta más natural al momento de estimar θ_k:

$$X_k = \theta_k + \epsilon\sigma_k\xi_k, \qquad k = 1, 2, \ldots \tag{2.1.28}$$

donde la nueva "data" viene dada por $X_k = y_k/b_k$, y $\sigma_k = b_k^{-1} > 0$. La dificultad de un problema lineal inverso puede ser definida por el comportamiento de σ_k: así, cuando $\sigma_k \to \infty$ a medida que $k \to \infty$, el problema es mal puesto, pero el tipo de crecimiento de σ_k puede ser usado para definir una medida de la mala postura. De ahí, la siguiente definición:

DEFINICIÓN 2.1.4. Un problema inverso se dice que es levemente mal posicionado si la sucesión σ_k tiene un crecimiento polinomial ($\sigma_k \approx k^\beta$) a medida que $k \to \infty$ y rigurosamente mal puesto si σ_k se incrementa con razón exponencial ($\sigma_k \approx \exp(\beta k)$) para algún $\beta > 0$. β es llamado el grado de mal posicionamiento del problema inverso. [1]

La teoría que se irá desarrollando en este trabajo sobre los problemas inversos levemente mal condicionados, haciendo la acotación cuando se crea oportuno.

§2.1.2. Estimación no paramétrica

Aproximación Minimax

Se tiene el modelo $X_k = \theta_k + \epsilon\sigma_k\xi_k$, con $k = 1, 2, \ldots$; Sea $\widehat{\theta} = \widehat{\theta}(X) = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots)$ un estimador de $\theta = (\theta_1, \theta_2, \ldots)$ basado en la data $X = \{X_k\}$. Recordemos que f se puede escribir como: $f = \sum_k \theta_k\varphi_k = \sum_k \langle f, \varphi_k\rangle\varphi_k$, donde $\{\varphi_k\}$ es la base ortogonal de vectores propios del operador A. Así, un estimador para f es $\widehat{f} = \sum_k \widehat{\theta}_k\varphi_k$. La pregunta que ahora nos hacemos es que tan bueno es este estimador.

Cuando se habla de un estimador hablamos de un elemento que es aleatorio, si queremos ver de forma cuantitativa que tan bueno es $\widehat{f}$ como estimador de f, calculamos la esperanza del cuadrado de la diferencia entre $\widehat{f}$ y f. Definimos la función riesgo como la media integrada del error al cuadrado (mean integrated squared error(MISE)), de $\widehat{f}$ como:

$$\mathcal{R}(\widehat{f}, f) = E_f\|\widehat{f} - f\|^2, \tag{2.1.29}$$

veamos que resulta del lado derecho.

$$\begin{aligned}
\|\widehat{f} - f\|^2 &= \|\sum_{k=1}^{\infty}\widehat{\theta}_k\varphi_k - \sum_{k=1}^{\infty}\theta_k\varphi_k\|^2 && (2.1.30)\\
&= \|\sum_{k=1}^{\infty}\varphi_k(\widehat{\theta}_k - \theta_k)\|^2 && (2.1.31)\\
&= \sum_{k=1}^{\infty}\varphi_k(\widehat{\theta}_k - \theta_k)\cdot\sum_{k=1}^{\infty}\varphi_k(\widehat{\theta}_k - \theta_k) && (2.1.32)\\
&= \sum_{k=1}^{\infty}(\widehat{\theta}_k - \theta_k)^2 \quad (\{\varphi_k\} \text{ base ortonormal}) && (2.1.33)\\
&= \|\widehat{\theta} - \theta\|^2 && (2.1.34)
\end{aligned}$$

Así,

$$\mathcal{R}(\widehat{f}, f) = E_f\|\widehat{f} - f\|^2 = E_\theta\|\widehat{\theta} - \theta\|^2 \tag{2.1.35}$$

donde la notación $\|\cdot\|$ es para la norma $l^2 - norma$ o $\|\cdot\|_2$ cuando es aplicada a los $\theta-$vectores en el espacio de sucesiones. De aquí en adelante E_f y E_θ denotarán la esperanza con respecto a Y o $X = (X_1, X_2, \ldots)$ para los modelos (2.1.2) y (2.1.28) respectivamente. Entonces (2.1.35) nos dice que analizar la función riesgo $\mathcal{R}(\widehat{f}, f)$ del es-

timador $\widehat{f}$ es igual a analizar el valor esperado $E_\theta = \|\widehat{\theta} - \theta\|^2$ en el espacio de sucesiones.

Acá se presenta una limitación al momento de minimizar la media integrada del error al cuadrado, porque como se puede ver la expresión depende de f y θ, las cuales son desconocidas.

Así, para buscar el estimador que minimiza se considerará una medida del riesgo tal como el riesgo minimax.

DEFINICIÓN 2.1.5. Supongamos que conocemos *a priori* la clase de funciones $\mathcal{F}$ a la que pertenece f. El riesgo minimax sobre $\mathcal{F}$ es definido como

$$r_\epsilon(\mathcal{F}) = \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathcal{R}(\widehat{f}, f) \tag{2.1.36}$$

donde el ínfimo es determinado sobre el conjunto de todos los estimadores de f. [1]

Dado a que los requerimientos para f se limitan sólo a que esta pertenece a una clase de funciones (suaves por ejemplo), será común que no se encuentren estimadores que alcancen el riesgo minimax, dado a que se podría tener un universo muy amplio de funciones en la clase considerada.

Consideremos por tanto, propiedades asintóticas como el nivel de ruido tendiendo a cero ($\epsilon \to 0$). Como ya se mencionó, en el modelo Gaussiano de ruido blanco esta tendencia es equivalente a considerar $n \to \infty$ en el modelo discreto.

Es claro que para cualquier estimador f^*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathcal{R}(\widehat{f}, f) \leqslant \sup_{f \in \mathcal{F}} \mathcal{R}(f^*, f) \tag{2.1.37}$$

De la desigualdad anterior surge al inquietud por definir estimador óptimo.

DEFINICIÓN 2.1.6. Un **estimador** f^* se dice **óptimo** ó que alcanza razón óptima de convergencia v_ϵ si la sucesión positiva v_ϵ converge a cero conforme $\epsilon \to 0$ y existen constantes C_1, C_2 satisfaciendo $0 < C_2 \leqslant C_1 < \infty$ tal que

$$C_2 v_\epsilon \leqslant \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathcal{R}(\widehat{f}, f) \leqslant \sup_{f \in \mathcal{F}} \mathcal{R}(f^*, f) \leqslant C_1 v_\epsilon \tag{2.1.38}$$

con $\epsilon \to 0$.

El estimador f^* se dice que es minimax o que es exacto si $C_1 = C_2$. [1]

De ahí que decimos que un estimador óptimo es uno cuyo riesgo es comparable al del mejor posible estimador.

§2.1.3. Clases de funciones

La intención ahora es definir clases de funciones que suministren información previa sobre la función (por ejemplo suavidad), dicha información puede ser usada para determinar características o propiedades de los coeficientes $\{\theta_k\}$.

Si suponemos f en una clase de funciones que corresponda a los elipsoides Θ en el espacio de los coeficientes $\{\theta_k\}$

$$\Theta = \Theta(a, L) = \{\theta : \sum_{k=1}^{\infty} a_k^2 \theta_k^2 \leqslant L\}, \tag{2.1.39}$$

donde $a = \{a_k\}$ es una sucesión no negativa y $a_k \to \infty$ cuando $k \to \infty$, $L > 0$. Así, cuando $k \to \infty$ la sucesión $\{\theta_k\}$ ha de ser decreciente, tomando valores pequeños para k grande. [5]

Los elipsoides resultan ser la forma más natural de obtener un comportamiento de tipo ℓ_2 de los coeficientes; se tomará en cuenta cuando se mencione la aproximación minimax adaptada.

§2.1.4. Métodos de Regularización

Distinto es el significado que se tiene del término regularización dependiendo del contexto en el que se trabaja, pero la esencia es la misma. Por ejemplo, en problemas inversos cuando hablamos de regularización, nos referimos a los mecanismos que utilizamos para obtener las soluciones de un problema inverso mal puesto. En el mundo de la estadística para obtener los estimadores que resuelven problemas con incertidumbre de fiabilidad e inestabilidad. En proceso de imágenes, regularización es sinónimo de reconstrucción de imágenes.

Los métodos de regularización serán definidos sobre el dominio espectral del operador, debido a que propiedades teóricas de dichos métodos pueden ser estudiados más fácilmente sobre el espectro. [1]

Recordemos el modelo $X_k = \theta_k + \epsilon\sigma_k\xi_k$, con $k = 1, 2, \ldots$, donde X_k representa la data por nosotros medida y $\sigma_k = b_k^{-1}$ (b_k los valores singulares de la SVD). En este modelo se desea estimar el parametro θ.

Consideremos una sucesión $\lambda = (\lambda_1, \lambda_2, \ldots)$ de valores ponderados no aleatorios, a través del cual se estima θ de la siguiente manera: $\hat{\theta}_k = \lambda_k X_k$. Además,

$$\hat{f} = \sum_{k=1}^{\infty} \hat{\theta}_k(\lambda_k)\varphi_k \tag{2.1.40}$$

Totalmente análogo al caso de dimensión finita considerado en [6]

Ahora consideraremos la descomposición del valor singular truncado (TSVD) escrita en términos de la función indicatriz. Nuestro parámetro de regularización ahora es un entero N, considerando así, $\lambda_k = I(k \leqslant N)$. Donde I denota la función indicatriz,la cual viene dada como:

$$\lambda_k = \begin{cases} 1, & \text{si } k \leqslant N; \\ 0, & \text{si } k > N; \end{cases} \tag{2.1.41}$$

Estos estimadores λ_k son llamados estimadores proyección. Así, se tiene

$$\widehat{\theta}_k(N) = \begin{cases} X_k, & \text{si } k \leqslant N; \\ 0, & \text{si } k > N; \end{cases} \tag{2.1.42}$$

N es conocido como ancho de banda.
La TSVD en el caso finito dimensional es presentada en [1] y se hizo un esbozo de la misma en el capítulo anterior.

Una característica poco favorable a este método es que la estimación del parámetro es muy brusca, esto debido a la rigidez de los valores que puede arrojar (0 o X_k), faltando suavidad en la estimación por decirlo de alguna manera. Además, no es eficiente computacionalmente, ya que requiere el cálculo de la SVD.

Otro método de regularización bastante conocido es el de Tikhonov. El propósito es encontrar un estimador que minimice el funcional

$$\min_g \{\|Ag - y\|^2 + \gamma\|g\|^2\} \tag{2.1.43}$$

donde $\gamma > 0$ es el parámetro de regularización.

Bajo condiciones de regularidad, el minimizador de (2.1.43) es $\widehat{f}_\gamma = (A^*A + \gamma I)^{-1}A^*Y$. Estos estimadores también pueden ser definidos en el dominio espectral usando los ponderados $\lambda_k = \frac{1}{1+\gamma\sigma_k^2}$ con $\gamma > 0$. Entonces el estimador es definido como en (2.1.40).

Una observación que se puede hacer de la regularización de Tickhonov, es que existen otras generalizaciones de esta, esto respecto al funcional g que se considera. Por ejemplo, se puede citar la función definida Ω definida en el capítulo anterior cuando se describió la regularización.

Estudiemos una nueva función que nos proporcione información de que tan bueno es un estimador.

La función L^2-riesgo del estimador lineal dado en (2.1.40) viene dada por:

$$\begin{aligned}
\mathcal{R}(\widehat{f}(\lambda), f) &= R(\theta, \lambda) && (2.1.44)\\
&= E_\theta \sum_k (\widehat{\theta}_k(\lambda_k) - \theta_k)^2 \qquad \textit{por (2.1.29, 2.1.33)} && (2.1.45)\\
&= E_\theta \sum_k (\lambda_k X_k - \theta_k)^2 \qquad \textit{por (2.1.40)} && (2.1.46)\\
&= E_\theta \sum_k (\lambda_k\theta_k + \epsilon\sigma_k\lambda_k\xi_k - \theta_k)^2 \qquad \textit{por (2.1.28)} && (2.1.47)\\
&= E_\theta \sum_k ((\lambda_k - 1)\theta_k + \epsilon\sigma_k\lambda_k\xi_k)^2 && (2.1.48)\\
&= E_\theta \sum_k ((\lambda_k - 1)^2\theta_k^2 + 2(\lambda_k - 1)\theta_k\epsilon\sigma_k\lambda_k\xi_k + \epsilon^2\sigma_k^2\lambda_k^2\xi_k^2) && (2.1.49)\\
&= \sum (E_\theta(\lambda_k - 1)^2\theta_k^2 + 2(\lambda_k - 1)\theta_k\epsilon\sigma_k\lambda_k E_\theta(\xi_k) + \epsilon^2\sigma_k^2\lambda_k^2 E_\theta(\xi_k^2)) && (2.1.50)
\end{aligned}$$

Ya que ξ es ruido blanco, $\{\xi_k\}$ es una sucesión de variables aleatorias Gaussianas estandar identicamente distribuidas, es decir, con distribución $\mathcal{N}(0, 1)$. De ahí que $E_\theta(\xi_k) = 0$ y $E_\theta(\xi_k^2) = 1$. Entonces,

$$\mathcal{R}(\widehat{f}(\lambda), f) = \sum_{k=1}^{\infty}(\lambda_k - 1)^2\theta_k^2 + \epsilon^2\sum_{k=1}^{\infty}\sigma_k^2\lambda_k^2 \qquad (2.1.51)$$

Donde la primera suma en (2.1.51) denota el sesgo y la segunda denota la varianza [1]. El sesgo proporciona información de si el estimador es una buena aproximación de la función desconocida f sin la presencia de ruido. La varianza medirá la variabilidad del estimado de inversión introducido por el ruido aleatorio [1]. Sesgo y varianza deberian ser pequeños para garantizar buena estimación.

Por simplicidad se considerarán los estimadores proyección para los métodos que ahora se abordan. Así, iniciamos con el riesgo para estos estimadores con ancho de banda N, el cual de la expresión (2.1.51) resulta:

$$R(\theta, N) = \sum_{k=N+1}^{\infty} \theta_k^2 + \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 \tag{2.1.52}$$

Notese la fuerte dependencia hacia N del resultado anterior; determinar este N es un importante problema en la estadística no-paramétrica. Debido a que el sesgo depende del parámetro desconocido f, se hace necesario obtener una cota para la función riesgo. Dicha cota nos la da el siguiente

TEOREMA 2.1.1. *Considere el caso donde $\sigma_k = k^\beta$ (problema inverso levemente mal puesto) con $\beta > 0$ y $\theta \in \Theta(\alpha, L)$, donde $a_k = k^\alpha$. Entonces el estimador dado como en (2.1.42) con $N \sim \epsilon^{\frac{-2}{2\alpha+2\beta+1}}$ cuando $\epsilon \to 0$ verifica*

$$\sup_{\theta \in \Theta(\alpha,L)} R(\theta, N) \leqslant C\epsilon^{\frac{4\alpha}{2\alpha+2\beta+1}} \tag{2.1.53}$$

Prueba

Por (2.1.52),

$$\sup_{\theta \in \Theta(\alpha,L)} R(\theta, N) = \sup_{\theta \in \Theta(\alpha,L)} \sum_{k=N+1}^{\infty} \theta_k^2 + \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 \tag{2.1.54}$$

Además,

$$\sup_{\theta \in \Theta(\alpha,L)} \sum_{k=N+1}^{\infty} \theta_k^2 = \sup_{\theta \in \Theta(\alpha,L)} \sum_{k=N+1}^{\infty} k^{2\alpha} \theta_k^2 k^{-2\alpha} \tag{2.1.55}$$

$$\sup_{\theta\in\Theta(\alpha,L)}\sum_{k=N+1}^{\infty}\theta_k^2 = \sup_{\theta\in\Theta(\alpha,L)}\sum_{k=N+1}^{\infty}k^{2\alpha}\theta_k^2k^{-2\alpha} \tag{2.1.56}$$

$$\leqslant \sup_{\theta\in\Theta(\alpha,L)}\sum_{k=N+1}^{\infty}k^{2\alpha}\theta_k^2N^{-2\alpha} \qquad (\alpha>0 \text{ y } k>N) \tag{2.1.57}$$

$$= N^{-2\alpha}\sup_{\theta\in\Theta(\alpha,L)}\sum_{k=N+1}^{\infty}(k^{\alpha})^2\theta_k^2 \tag{2.1.58}$$

$$= LN^{-2\alpha} \qquad (\theta\in\Theta(\alpha,L) \text{ con } a_k=k^{\alpha}) \tag{2.1.59}$$

Por otro lado,

$$\epsilon^2\sum_{k=1}^{N}\sigma_k^2 \approx \epsilon^2\frac{N^{2\beta+1}}{2\beta+1} \quad \text{para } N \text{ suficientemente grande} \tag{2.1.60}$$

Por tanto,

$$\sup_{\theta\in\Theta(\alpha,L)}R(\theta,N)\leqslant LN^{-2\alpha}+\epsilon^2\frac{N^{2\beta+1}}{2\beta+1} \tag{2.1.61}$$

Si se tiene $N\sim\epsilon^{\frac{-2}{2\alpha+2\beta+1}}$ y $\epsilon\to 0$ se concluye que:

$$\sup_{\theta\in\Theta(\alpha,L)}R(\theta,N)\leqslant L\epsilon^{\frac{4\alpha}{2\alpha+2\beta+1}} \tag{2.1.62}$$

Las nociones de adaptación y oráculos intentan responder a la forma de como tomar el ancho de banda sin tener que hacer fuertes suposiciones sobre la función f.

§2.2. Adaptación y desigualdades oráculos

La calibración de los estimadores dentro de una clase específica, donde por ejemplo ya hemos considerado en los métodos de regularización N para el $TSVD$ o γ para Tickonov, es un problema de mucha importancia dentro de la estadística no-paramétrica. Es importante notar que el teorema (2.1.1) nos indica la dependencia que tiene una óptima razón de convergencia respecto a la toma del parámetro de regularización en nuestro caso.

§2.2.1. Estimación Minimax Adaptada

Se considera una colección $\mathcal{A} = \{\Theta_\alpha\}$ de clases $\Theta_\alpha \subset \ell_2$. En la práctica se sabe que θ pertenece a alguna de las clases de $\mathcal{A}$, pero *a priori*, no se cuenta con la información de a cual de las clases pertenece. Cuando Θ_α es una clase cuyos elementos presenta suavidad, lo anterior puede describirse como el hecho que conocemos que nuestra función es suave pero se desconoce "que tan suave" es.

La noción de estimación minimax adaptada ha sido desarrollada para definir estimadores que "se adaptan" a la suavidad desconocida de la función.

DEFINICIÓN 2.2.1. Un estimador $\theta^\star$ es llamado minimax adaptado sobre la escala de clases de $\mathcal{A}$ si para cada $\Theta_\alpha \in \mathcal{A}$ el estimador $\theta^\star$ consigue la razón óptima de convergencia.

De la definición anterior se deriva la importancia que tienen los estimadores minimax adaptados, debido a que estos son óptimos para cualquier parámetro α en la colección $\mathcal{A}$. La adaptividad garantiza una buena precisión para una gran gama de funciones, el estimador correspondiente se "adapta" a la suavidad de la función en estudio.

§2.2.2. Desigualdades Oráculo

En la estimación minimax adaptada, el estimador óptimo es tomado de una colección de clases de estimadores. Ahora desde la aproximación de desigualdades oráculo se fija la clase de estimadores y se busca el mejor estimador dentro de esa clase. Un tratamiento parecido se tiene en la aproximación minimax, donde el objetivo es el de obtener una función $\widehat{f}$ dentro de una clase dada que sea el mejor estimador de f, esto se lleva a cabo buscando cota para la función riesgo.

Consideremos el estimador $\widehat{\theta}$ con las características como en (2.1.40); este se obtiene a través de unos valores ponderados λ_i. Así, podríamos considerar la clase de los pesos λ que definen $\widehat{\theta}$ como la clase Λ.

Por otra parte, definamos el oráculo λ^0 de tal forma que se satisfaga

$$R(\theta, \lambda^0) = \inf_{\lambda \in \Lambda} R(\theta, \lambda) \tag{2.2.1}$$

Dicho de otra forma, λ^0 es el que minimiza la función riesgo. Notese la dependencia de (2.2.1) respecto a θ, es decir, (2.2.1) tiene sentido si se conoce θ, pero ya que θ es desconocido para nosotros, simplemente el λ^0 es una predicción del mismo y por ello llamado oráculo.

Se desea trabajar con una sucesión $\lambda^\star$ que dependa de los datos y que siga tomando valores en Λ de manera que el estimador $\theta^\star = \widehat{\theta}(\lambda^\star)$ satisfaga $\forall\theta \in \ell^2$:

$$E_\theta\|\theta^* - \theta\|^2 \leqslant C_\epsilon \inf_{\lambda\in\Lambda} R(\theta,\lambda) + \Delta_\epsilon \tag{2.2.2}$$

La desigualdad anterior se le llama desigualdad oráculo, en dicha desigualdad C_ϵ y Δ_ϵ son términos dependientes de un valor ϵ tendiendo a cero. Un par de ejemplos de estos valores se tendrán para dos desigualdades oráculos que se mostrarán más adelante. Además, si Δ_ϵ es pequeño (es decir, menor que $R(\theta,\lambda^0)$) entonces una desigualdad oráculo garantiza que el estimador tiene un riesgo del mismo orden que el oráculo. La intención ahora es encontrar métodos que dependan de la data y nos arrojen una selección automática de λ o al menos se parezcan mucho al oráculo.

Se pueden presentar resultados asintóticos:

DEFINICIÓN 2.2.2. Un estimador θ^* se dice que satisface una desigualdad oráculo exacta sobre la clase Λ conforme $\epsilon \to 0$ si

$$E_\theta\|\theta^* - \theta\|^2 \leqslant (1 + o(1))R(\theta,\lambda^0) \tag{2.2.3}$$

para cada θ en algún subconjunto $\Theta_0 \subseteq \ell^2$

Un estimador como el de la definición anterior es bastante parecido al oráculo sobre Λ para cualquier sucesión $\theta \in \Theta_0$.

§2.2.3. Selección de modelo y estimación de riesgo

Aquí se considerará la selección de modelo como el problema de tomar el mejor estimador de una familia Λ; esta selección se hará dependiendo de la data.

Estimación del riesgo insesgado

El parámetro θ es desconocido, y de ahí que la función riesgo también lo sea. Por tanto, se entiende que se busque para un estimador que minimice un estimado del

riesgo basado en la data.

Una aproximación a este problema de minimización se basa en el principio de estimación de riesgo insesgado (URE). Esto en el contexto de estimadores lineales $\widehat{\theta}(\lambda)$ se considera el estimador de riesgo de Stein dado como

$$U(X,\lambda) = \sum_{k=1}^{\infty}(1-\lambda_k)^2(X_k^2 - \epsilon^2\sigma_k^2) + \epsilon^2\sum_{k=1}^{\infty}\sigma_k^2\lambda_k^2. \tag{2.2.4}$$

$U(X,\lambda)$ es un estimador insesgado de $R(\theta,\lambda)$. En efecto,

$$\begin{aligned}
E_\theta U(X,\lambda) &= \sum_{k=1}^{\infty}(1-\lambda_k)^2 E_\theta(X_k^2 - \epsilon^2\sigma_k^2) + \epsilon^2\sum_{k=1}^{\infty}\sigma_k^2\lambda_k^2 && (2.2.5)\\
&= \sum_{k=1}^{\infty}(1-\lambda_k)^2[E_\theta(\theta_k^2) + 2\theta_k\epsilon\sigma_k E_\theta(\xi_k) + \epsilon^2\sigma_k^2 E_\theta(\xi_k^2) - \epsilon^2\sigma_k^2] + \epsilon^2\sum_{k=1}^{\infty}\sigma_k^2\lambda_k^2 &&\\
&= \sum_{k=1}^{\infty}(1-\lambda_k)^2[(\theta_k^2) + +\epsilon^2\sigma_k^2 - \epsilon^2\sigma_k^2] + \epsilon^2\sum_{k=1}^{\infty}\sigma_k^2\lambda_k^2 && (2.2.6)\\
&= \sum_{k=1}^{\infty}(1-\lambda_k)^2\theta_k^2 + \epsilon^2\sum_{k=1}^{\infty}\sigma_k^2\lambda_k^2 && (2.2.7)\\
&= R(\theta,\lambda) \qquad \text{por (2.1.51)} && (2.2.8)
\end{aligned}$$

$\therefore U(X,\lambda)$ es un estimador insesgado de $R(\theta,\lambda)$ para todo λ. Lo anterior es equivalente a considerar solo $X_k^2 - \epsilon^2\sigma_k^2$ como un estimador insesgado de θ_k^2 en (2.1.51).

El principio de la estimación del riesgo insesgado es entonces encontrar el λ que minimiza el riesgo estimado $U(X,\lambda)$:

$$\lambda^* = \arg\min_{\lambda\in\Lambda} U(X,\lambda). \tag{2.2.9}$$

Ahora se dará un ejemplo de desigualdad oráculo obtenida cuando la estimación de riesgo insesgado es usada para estimar N (estimadores de proyección) ó γ para regularización de Tikhonov. Se define:

$$S = \left(\frac{\max_{\lambda\in\Lambda}\sum_{k=1}^{\infty}\sigma_k^4\lambda_k^2}{\min_{\lambda\in\Lambda}\sum_{k=1}^{\infty}\sigma_k^4\lambda_k^2}\right)^{1/2}. \tag{2.2.10}$$

El siguiente resultado citado en [1] es un ejemplo de una desigualdad oráculo que se obtiene a partir de la estimación del riesgo insesgado cuando se usa para obtener N (estimadores de proyección) o γ para regularización de Tickhonov.

TEOREMA 2.2.1. *Supongamos que $\sigma_k = k^\beta$, Λ es finito con cardinalidad D y el estimador es $\theta^* = (\theta_1^*, \theta_2^*, \ldots)$ con $\theta_k^* = \lambda_k^* X_k$. Entonces existen constantes $\gamma_1, \gamma_2 > 0$ tales que*

$$E_\theta \|\theta^* - \theta\|^2 \leqslant (1 + \gamma_1 B^{-1}) \min_{\lambda \in \Lambda} R(\theta, \lambda) + \gamma_2 B \epsilon^2 [\log(DS)]^{2\beta+1} \tag{2.2.11}$$

para cada $\theta \in \ell_2$ y para cada $B > 0$ suficientemente grande. [2]

En realidad al imponer restricciones débiles al comportamiento de D y S para valores grandes de ϵ, se puede obtener una desigualdad oráculo exacta. Esto significa, por ejemplo, que podemos imitar el método de Tikhonov con la mejor toma posible del parametro de calibración γ.

La demostración del teorema anterior implica todo el desarrollo de [2].

§2.2.4. Método de la cápsula de riesgo

Aún cuando se pueden obtener desigualdades oráculos precisas como lo presenta el teorema (2.2.1), para algunas muestras el estimador puede no ser preciso del todo, siendo esto mostrado por Cavalier al desarrollar [3]. Así, el riesgo medio sobre todas las muestras es muy grande. De ahí, a que se vaya en busqueda de selección del parámetro, que dependan de la data, más estables. Para esta parte se considerará sólo la clase de estimadores proyección.

El estimador de riesgo insesgado para estimadores de proyección es:

$$U(X, N) = \sum_{k=N+1}^{\infty} (X_k^2 - \epsilon^2 \sigma_k^2) + \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 \tag{2.2.12}$$

Minimizar $U(X, N)$ sobre N es equivalente a minimizar

$$\overline{R}(X, N) = \sum_{k=N+1}^{\infty} (X_k^2 - \epsilon^2 \sigma_k^2) + \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 - \|X - \epsilon\sigma\|^2 \tag{2.2.13}$$

sobre N, dado que el último término de la expresión no depende de N.

Ahora bien, como,

$$\|X - \epsilon\sigma\|^2 = \sum_{k=1}^{\infty} X_k^2 - 2\epsilon\langle X, \sigma\rangle + \epsilon^2 \sum_{k=1}^{\infty} \sigma_k^2,$$

nos queda:

$$\begin{aligned}
\overline{R}(X, N) &= \sum_{k=N+1}^{\infty} (X_k^2 - \epsilon^2\sigma_k^2) + \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 - \sum_{k=1}^{\infty} X_k^2 + 2\epsilon\langle X, \sigma\rangle - \epsilon^2 \sum_{k=1}^{\infty} \sigma_k^2 && (2.2.14)\\
&= \sum_{k=N+1}^{\infty} X_k^2 - \epsilon^2 \sum_{k=N+1}^{\infty} \sigma_k^2 + \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 - \sum_{k=1}^{\infty} X_k^2 + 2\epsilon\langle X, \sigma\rangle - \epsilon^2 \sum_{k=1}^{\infty} \sigma_k^2 && (2.2.15)\\
&= \sum_{k=N+1}^{\infty} X_k^2 - \epsilon^2 \sum_{k=N+1}^{\infty} \sigma_k^2 + \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 - \sum_{k=1}^{\infty} X_k^2 + 2\epsilon\langle X, \sigma\rangle - \epsilon^2 \sum_{k=1}^{\infty} \sigma_k^2 && (2.2.16)\\
&= \sum_{k=N+1}^{\infty} X_k^2 - \sum_{k=1}^{N} X_k^2 - \sum_{k=N+1}^{\infty} X_k^2 - \epsilon^2 \sum_{k=N+1}^{\infty} \sigma_k^2 + \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 - \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 - \epsilon^2 \sum_{k=N+1}^{\infty} \sigma_k^2 + 2\epsilon\langle X, \sigma\rangle &&\\
&= -\sum_{k=1}^{N} X_k^2 - 2\epsilon^2 \sum_{k=N+1}^{\infty} \sigma_k^2 + 2\epsilon\langle X, \sigma\rangle && (2.2.17)
\end{aligned}$$

de donde,

$$\overline{R}(X, N) = -\sum_{k=1}^{N} X_k^2 - 2\epsilon^2 \sum_{k=N+1}^{\infty} \sigma_k^2 + 2\epsilon\langle X, \sigma\rangle \qquad (2.2.18)$$

El último término de la última expresión no depende de N, por lo que minimizar $\overline{R}(X, N)$ respecto a N, es equivalente a minimizar

$$\overline{R}(X, N) = -\sum_{k=1}^{N} X_k^2 - 2\epsilon^2 \sum_{k=N+1}^{\infty} \sigma_k^2 \qquad (2.2.19)$$

El método de riesgo empírico penalizado es una aproximación general que es bastante similar a la del riesgo insesgado. Se comienza definiendo una versión "penalizada" de (2.2.19):

$$\overline{R}_{pen}(X, N) = -\sum_{k=N+1}^{N} X_k^2 - \epsilon^2 \sum_{k=N+1}^{\infty} \sigma_k^2 + pen(N), \qquad (2.2.20)$$

donde $pen(N)$ es una función de penalización. El ancho de banda tomado es entonces definido como:

$$N(X) = \arg\min_{N \geqslant 1} \overline{R}_{pen}(X, N). \tag{2.2.21}$$

De ahí que se considere a la estimación de riesgo insesgado como un riesgo empírico penalizado con una función de penalización específica, dada como

$$pen_{ure}(N) = -\epsilon^2 \sum_{k=N+1}^{\infty} \sigma_k^2. \tag{2.2.22}$$

La idea principal en la aproximación penalizada es que penalidades rigurosas logren resultados por encima de la estimación del riesgo insesgado. Aquí, la selección de la función de penalidad es crucial; así, se plantea un nuevo método, conocido como minimización de la cápsula de riesgo (risk hull minimization(RHM)); el cual mejora la (URE) dado que provee una buena estrategia para seleccionar la función de penalidad.

La motivación heuristica de la aproximación RHM es fundamentada en la idea de los oráculos. Partamos de tener un oráculo que nos da $\theta_k, k = 1, 2, \ldots$ y además usemos estimadores de proyección. Acá, se tendrá que el ancho de banda que minimiza es dado por $N_{or} = \arg\min_N r(X, N)$; donde,

$$r(X, N) = \|\widehat{\theta}(N) - \theta\|^2 = \sum_{k=N+1}^{\infty} \theta_k^2 + \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 \xi_k^2. \tag{2.2.23}$$

Ahora se intentará asemejar está selección del ancho de banda con un procedimiento dependiente de la data. El problema pareciera no soluble, dado al desconocimiento tanto de θ_k^2 como de ξ_k^2. Pero, supongamos que conocemos cada θ_k^2 y con esto buscamos minimizar $r(X, N)$. Ya que para nosotros ξ_k^2 es desconocido, se minimizará ahora el siguiente funcional no aleatorio:

$$l(\theta, N) = \sum_{k=N+1}^{\infty} \theta_k^2 + V(N), \tag{2.2.24}$$

donde $V(N)$ es una cota superior del término estocástico $\epsilon^2 \sum_{k=1}^{N} \sigma_k^2 \xi_k^2$. Naturalmente buscamos tomar $V(N)$ tal que

$$E \sup_N \left[\epsilon^2 \sum_{k=1}^{N} \sigma_k^2 \xi_k^2 - V(N) \right] \leqslant 0, \tag{2.2.25}$$

porque el riesgo para cualquier estimador proyección, con cualquier ancho de banda $\tilde{N}$ controlado por la data puede ser fácilmente controlado:

$$E_\theta \|\widehat{\theta}(\tilde{N}) - \theta\|^2 \leqslant E_\theta l(\theta, \tilde{N}). \tag{2.2.26}$$

El argumento recien descrito conduce a la siguiente

DEFINICIÓN 2.2.3. Una función no aleatoria $l(\theta, N)$ es llamada cápsula de riesgo si

$$E_\theta \sup_N [r(X, N) - l(\theta, N)] \leqslant 0. \tag{2.2.27}$$

De ahí que, $l(\theta, N)$ como en (2.2.24) y (2.2.25) es una cápsula de riesgo. Evidentemente, la cota superior (2.2.26) ha de ser lo más pequeña posible, de esta forma tenemos la cápsula minimal (notar la fuerte dependencia de esta cápsula sobre σ_k^2).

Una vez una función $V(N)$ satisfaciendo (2.2.25) ha sido tomada, la minimización de $l(\theta, N)$ puede hacerse usando estimación insesgada estandar. El problema se reduce a la minimización de $-\sum_{k=1}^{N} \theta_k^2 + V(N)$. Como antes, por reemplazar el parámetro desconocido θ_k^2 por su estimado insesgado $X_k^2 - \epsilon^2\sigma_k^2$, se llega a la siguiente selección de ancho de banda adaptativa:

$$\tilde{N} = \arg\min_N \left[-\sum_{k=1}^{N} X_k^2 + \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 + V(N) \right]. \tag{2.2.28}$$

En el marco de la minimización de riesgo empírico, el método de minimización de la cápsula de riesgo puede ser definida como sigue. Sea la función de penalidad en (2.2.21) definida como:

$$pen(N) = pen_{rhm}(N) = \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 + (1 + \alpha) U_0(N), \tag{2.2.29}$$

donde $\alpha > 0$ es una constante fija y

$$U_0(N) = \text{ínf}\{t > 0 : E(\eta_N I(\eta_N \geqslant t)) \leqslant \epsilon^2\sigma_1^2\}, \tag{2.2.30}$$

con $\eta_N = \epsilon^2 \sum_{k=1}^{N} \sigma_k^2(\xi_k^2 - 1)$. La función $U_0(N)$ puede ser calculada por simulaciones de Monte Carlo. El método de la cápsula de riesgo(RHM) toma los ancho de banda N_{rhm} de acuerdo a (2.2.21) con la función de penalidad definida por (2.2.29) y (2.2.30).

El RHM penalizado es el URE penalizado más el término $(1+\alpha)U_0(N)$. Cuando $N \to \infty$, se tiene

$$U_0(N) \approx \left(2\epsilon^4 \sum_{k=1}^{N} \sigma_k^4 log\left(\frac{\sum_{k=1}^{N} \sigma_k^4}{2\pi\sigma_1^4} \right) \right)^{\frac{1}{2}}. \tag{2.2.31}$$

La siguiente desigualdad oráculo nos da una cota superior para el riesgo al cuadrado medio de la aproximación RHM. Acá se asume que σ_k tiene crecimiento polinomial $\sigma_k = k^\beta$.

TEOREMA 2.2.2. *Sea la selección RHM del ancho de banda N_{rhm} como fué definida en (2.2.21) con la función de penalidad definida por (2.2.29) y (2.2.30) y θ^*_{rhm} el estimador proyección asociado. Entonces existen constantes $C_* > 0$ y $\delta_0 > 0$ tal que para todo $\delta \in (0, \delta_0]$ y $\alpha > 1$*

$$E\|\tilde{\theta}_{rhm} - \theta\|^2 \leqslant (1+\delta) \inf_N R_{rhm}(\theta, N) + C_* \epsilon^2 \left(\frac{1}{\delta^{4\beta+1}} + \frac{1}{\alpha - 1} \right). \tag{2.2.32}$$

donde

$$R_{rhm}(\theta, N) = \sum_{k=N+1}^{\infty} \theta_k^2 + \epsilon^2 \sum_{k=1}^{N} \sigma_k^2 + (1+\alpha)U_0(N) \tag{2.2.33}$$

El teorema anterior nos provee de una desigualdad oráculo pero con un término de penalidad sobre el riesgo en el lado derecho de la misma.

CONCLUSIONES

En este trabajo se describe la teoría de problemas inversos mal puestos desde la perspectiva del algebra lineal (Descomposición del valor singular) y desde la estimación de parámetros en estadística. Se hace una revisión del artículo [1] de Cavalier, donde se recopilan resultados acerca de la construcción de estimadores con regularización Tickonov en el ámbito de espacios de Hilbert.

En este estudio se revisaron las definiciones de problemas bien puestos y bien condicionados, a partir de las cuales se definió el problema inverso mal puesto y mal condicionado. Se desarrolló la teoría básica de problemas inversos, considerando el operador A como una transformación lineal (una matriz de orden $m \times n$). Para ello se recurrió a la descomposición del valor singular (SVD) de A, y a partir de allí, estudiar como es que A actúa sobre la imagen f en el modelo $Af = d$. Surge una dificultad al momento de reconstruir la imagen cuando se tienen valores singulares pequeños, debido a que estos amplifican el ruido a lo largo de las componentes de f donde estos influyen. [1].

Como alternativa se estudió los métodos de regularización, en particular, el método de Tickhonov y el de la descomposición del valor singular truncado. El primero busca una ponderación entre la información que puede ser sacada de la data y la que se pudiese obtener de información histórica sobre las cantidades a reconstruir. El segundo busca prescindir de la información de la imagen a lo largo de las componentes a partir de las cuales los valores singulares se hacen pequeños evitando la amplificación del ruido.

A partir de esta teoría, se busca una generalización de los problemas inversos, tratando estos en contextos más generales; es a partir de allí donde el problema $Af = d + n$, se comenzó a trabajar sobre el ámbito de los espacios de Hilbert. El operador envía la imagen f que ahora es una función en un espacio de Hilbert a otro espacio de Hilbert

donde se encuentra la data; además se añadió un nuevo factor: considerar el ruido una variable aleatoria, lo que permitió el estudio de los problemas inversos desde la estadística; en concreto, hablamos de estadística no paramétrica, porque temas teóricos básicos relacionados con esta, son aplicados a los problemas inversos mal puestos. Se trabajó con el modelo de ruido blanco y se usó la SVD del operador para discretizar el problema. A su vez la SVD nos provee forma de medir el mal posicionamiento del problema inverso.

Al definir estimadores usamos aproximación no paramétrica. El comportamiento de los estimadores es medido por una función riesgo, definida como la medida integrada del error cuadrado. Aquí, el cálculo de este riesgo es díficil por la dependencia que tiene el parámetro desconocido, esto motivó considerar el peor riesgo que se tiene cuando se pide a la función f estar en una clase específica de funciones e ir en busqueda de los estimadores que minimizan este riesgo, los que hemos definido como estimadores minimax. Pero en los problemas inversos no paramétricos, estimadores minimax pueden no existir. De ahí, que se estudió una desigualdad de optimalidad minimax asintótica a medida que el ruido va a cero. Por supuesto que estos resultados dependen de la clase de funciones sobre la que se trabaja.

También se abordó lo relacionado a la regularización, lo que permite definir buenos estimadores para problemas inversos mal puestos en el contexto más general que se ha ido desarrollando. El problema en los distintos métodos es la dependencia al parámetro de regularización respectivo.

Luego las desigualdades oráculo permiten obtener buenos estimadores para los métodos de regularización dando hipótesis a la función f. Se revisó lo concerniente a la minimización de un estimado del riesgo, presentando por ejemplo, el estimador de riesgo insesgado de Stein (URE). Para un estimador proyección, la idea de estimar el nivel de truncamiento vía minimización de la URE es un caso particular del método del riesgo empírico penalizado. Además, se presentó la minimización de la cápsula de riesgo como estrategia en la selección de la función de penalidad.

Entre los objetivos alcanzados, se puede decir que se logró un estudio de la teoría

básica de los problemas inversos, para luego entrar con un enfoque más general. Todo esto, acompañado de una revisión de conceptos y resultados básicos del Algebra lineal, el análisis funcional y la estadística. Por otra parte, como estudios futuros relacionados a este trabajo, se puede considerar la experimentación numérica de algunos métodos relacionados a los problemas inversos, aplicados a ejemplos que surgen de diversas áreas de las ciencias y la ingenieria como son presentados en [7]; así como el estudio teórico, citando por ejemplo los resultados sobre desigualdades oráculos presentados acá y que constituyen el desarrollo de [2].

REFERENCIAS

[1] Cavalier L. 2007. Nonparametric statistical inverse problems.

[2] Cavalier L., G. K. Golubev, D. Picard and A. B. Tsybakov. *Oracle Inequalities for Inverse Problems.*The Annals of Statistics 2002, Vol. 30, No. 3, 843-874.

[3] Cavalier L., G. K. Golubev. *Risk hull method and regularization by projections of ill-posed Inverse Problems.*The Annals of Statistics 2006, Vol. 34, 1653.

[4] Meyer Carl D.*Matrix Analysis and Applied Linear Algebra*. SIAM, 2000.

[5] [Eduard N. Belitser and Boris Y. Levit] On Minimax Filtering over Ellipsoids *Math. Meth. Statist. Mathematical Institute, University of Utrecht. Netherlands*, 1995.

[6] S.M. Tan and Colin Fox. Physics 707. Inverse Problems. The University of Auckland.

[7] Tarantola, A.: *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, SIAM, 2005.

[8] DeVito Carl L. *Functional Analysis and Linear Operator Theory*. Addison-Wesley Publishing Company 1990.