UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRADO DE MATEMÁTICAS

Trabajo fin de grado

LA PARADOJA DE STEIN

Autor: Clàudia Casanovas Pato

Director:	Dr. Josep Fortiana Gregori
Realizado en:	Departamento de Matemáticas e Informática
Barcelona,	20 de junio de 2021

A ti Ponxi.

Abstract

The (population) mean of a p-dimensional multivariate normal vector is plainly estimated by the empirical mean which, additionally, is minimax, ML, UMV and least squares BLUE. One would fancy it is also best as to risk. Nonetheless, Stein (1956) proved it is inadmissible for $p > 2$, showing alternative, better candidates. This is Stein's paradox, origin of this memoir.

We begin with a brief introduction to place Stein's result in its proper historical context. Then, after reviewing some basic Statistics concepts we present Stein's result, accompanied by illustrative simulations. Finally we survey several approaches to understanding the paradox.

Resumen

La media (poblacional) de un vector normal multivariante p-dimensional se estima claramente mediante la media empírica que, además, es un estimador minimax, MV, UMV y por mínimos cuadrados MELI. Uno podría imaginar que también es el mejor en cuanto a su riesgo. No obstante, para $p > 2$ Stein (1956) demostró que es inadmisible, proporcionando estimadores alternativos, con menor riesgo. Esta es la paradoja de Stein, origen de estas memorias.

Comenzamos con una breve introducción para ubicar el resultado de Stein en su contexto histórico adecuado. Luego, después de revisar algunos conceptos básicos de Estadística, presentamos el resultado de Stein, acompañado de simulaciones ilustrativas. Finalmente, examinamos varios enfoques para comprender la paradoja.

2010 Mathematics Subject Classification. 62-03, 62H12, 62F10, 62F12, 62F15, 62J07, 62F99

Agradecimientos

En primer lugar, quiero dar las gracias a mi familia, pareja y amigos, por apoyarme desde que empecé esta carrera y por creer en mí, incluso cuando yo no lo hacía. Sin vosotros no lo habría logrado.

También quisiera agradecer a mi tutor, Dr. Josep Fortiana Gregori, por guiarme en el tema de este trabajo y por todo el soporte y tiempo dedicados para que este proyecto pueda ver la luz.

Índice

1. El sentido común engaña

En 1956, Charles Stein publicó un artículo que cambió para siempre el enfoque estadístico de la estimación de altas dimensiones. Su descubrimiento de que el estimador habitual del vector de medias normales es dominado en dimensiones tres y superiores sorprendió a muchos en ese momento, y se convirtió en el catalizador de una vasta y rica literatura.

Desconcertante e inaudito. Esa fue la primera impresión de muchos, ya que Stein justificó combinar observaciones no relacionadas para estimar sus valores esperados.

Por ejemplo, resulta que la media de nacimientos en Cataluña el año 2000 y el número de espectadores en el mundial de fútbol de 2010 permiten mejorar la estimación del precio del quilo de patatas en China el 2019. ¿Cómo puede ocurrir esto?

Intuitivamente pensamos que la mejor manera de estimar el precio del quilo de las patatas es observando precisamente esta variable, en vez de incluir en el modelo otras variables no relacionadas. De ahí que nos extrañemos ante el enunciado y lo consideremos inicialmente como una afirmación contradictoria e ilógica. Esta es la paradoja de Stein que da título a este trabajo.

Este fenómeno ocurre porque el coste de una mala estimación en una componente del vector de medias se ve compensado por una estimación mejor en otra componente, es decir, su riesgo total disminuye.

2. Los años 1950

2.1. Trasfondo teórico

Pese a contar con conceptos como el teorema de Bayes, mínimos cuadrados, el teorema central del límite, regresiones y correlaciones, métodos binomiales y de Poisson, etcétera, no había una teoría que englobara toda la disciplina estadística.

R. A. Fisher (1922), *On the Mathematical Foundations of Theoretical Statistics* y (1925) *Theory of Statistical Estimation* fueron un punto de inflexión. Introdujo ideas tales como consistencia, suficiencia, eficiencia, información de Fisher o máxima verosimilitud.

Esta cohesión fue reforzada por la serie de trabajos de J. Neyman y E. Pearson que culminaron en el artículo *"IX. On the problem of the most efficient tests of statistical hypotheses"* (1933), en el que establecieron las bases de la teoría de tests de hipótesis. Y por Kolmogorov (1933) donde proporcionó una base axiomática para la teoría de probabilidad.

Abraham Wald en teoría de decisión (1939, 1940, 1945, 1947), y (1950) *Statistical Decision Functions*, donde introdujo el concepto de función de decisión, riesgo, admisibilidad, procesos minimax, etcétera.

Destacamos el tratado de Savage L.J. (1954), *Foundations of Statistics*.

Kendall (1957), Roy (1957) y Anderson (1958) escribieron sobre análisis multivariante, siendo el libro de éste último el más conocido por su uso como libro de texto.

Robbins (1956) introdujo el método de Bayes empírico, un procedimiento para "estimar" (hiper)parámetros de la distribución a priori de los parámetros de un modelo.

Tanto Hodges y Lehmann, como Girshick y Savage, y Blyth (1951) demostraron la admisibilidad para $p = 1$. Stein demostró la admisibilidad en dos dimensiones usando el mismo método que Lehmann y Hodges.

Teorema 2.1.1. *El estimador usual de la media de una variable aleatoria normal $p-$variante, $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$, con σ conocida, es admisible para $p \leq 2$.*

Demostración. Tomamos $\sigma^2 = 1$ sin pérdida de generalidad. Vemos el caso $p = 2$ ya que el $p = 1$ se desprende de la misma demostración.

Sea δ un estimador de θ con riesgo finito y s su sesgo, esto es,

$$s(\theta) \equiv \text{Sesgo}(\delta) = \mathbb{E}(\delta(X)) - \theta. \tag{2.1}$$

Entonces por la cota de Crámer-Rao para estimadores sesgados,

$$Var(\delta) \geq \frac{(1 + s'(\theta))^2}{\mathcal{I}((\theta)}.$$

Lo que equivale a

$$\mathcal{R}(\theta) = \mathbb{E}\left[\|\delta - \theta\|^2\right] \geq s^2(\theta) + \frac{(1 + s'(\theta))^2}{\mathcal{I}(\theta)}. \tag{2.2}$$

En nuestro caso p-variante, la información de Fisher $\mathcal{I}(\theta)$ es la matriz identidad $p \times p$.

De Hodges y Lehmann (1951) tenemos,

$$\mathcal{R}(\theta) \geq s^2(\theta) + \sum_i \left(\sum_j \eta_{ij}^2 \, [\gamma_{ij} + s_{ij}(\theta)] \right)^2. \tag{2.3}$$

Donde,

- η es tal que $\sum_j \eta_{ij}^2 = 1$ para todo i.
- $\gamma_{ij} = 1$ si $i = j$ y 0 de lo contrario.
- $s_{ij} = \dfrac{\partial}{\partial \theta_j} s_i(\theta)$, donde s_i es la coordenada i-ésima de $s(\theta)$.

Tomando

$$\eta_{ij} = \frac{\gamma_{ij} + s_{ij}(\theta)}{\sqrt{\sum_j [\gamma_{ij} + s_{ij}(\theta)]^2}}, \tag{2.4}$$

maximizamos el término de la derecha de 2.3 y obtenemos,

$$\begin{aligned} \mathcal{R}(\theta) &\geq s^2(\theta) + \sum_{i,j} [\gamma_{ij} + s_{ij}(\theta)]^2 \\ &= s^2(\theta) + p + 2 \sum_i s_{ii}(\theta) + \sum_{i,j} s_{ij}(\theta). \end{aligned} \tag{2.5}$$

Dado que podemos restringirnos al caso esféricamente simétrico (ver proposición 6.2.1), s es de la forma

$$s(\theta) = -\varphi(\|\theta\|^2)\theta, \tag{2.6}$$

donde φ es una función real diferenciable. De este modo, obtenemos

$$\mathcal{R}(\theta) \geq p + \|\theta\|^2 \varphi\varphi^2(\|\theta\|^2) - 2p\varphi(\|\theta\|^2) - 4\|\theta\|^2 \varphi'(\|\theta\|^2). \tag{2.7}$$

Por 6.2.1, si $\theta_0 = X$ es inadmisible existe un estimador esféricamente simétrico δ estrictamente mejor, y por lo tanto existe una función φ no idénticamente nula tal que

$$2 \geq \mathcal{R}(\delta) \geq 2 + \|\theta\|^2 \varphi^2(\|\theta\|^2) - 4\varphi(\theta) - 4\|\theta\|^2 \varphi'(\|\theta\|^2), \qquad \forall \|\theta\|^2 > 0. \tag{2.8}$$

Haciendo el cambio de variables $t = \|\theta\|^2$ y $\psi(t) = t\varphi(t)$, obtenemos

$$0 \geq \frac{1}{t}\psi^2(t) - 4\psi'(t), \qquad \forall t > 0. \tag{2.9}$$

Esto muestra que ψ es una función creciente. Mostraremos por reducción al absurdo que 2.9 implica que ψ es idénticamente 0. Suponemos primeramente que $\psi(t_0) < 0$ para algún $t_0 > 0$. Entonces integrando la desigualdad,

$$\frac{\psi'(t)}{\psi^2(t)} \geq \frac{1}{4t} \tag{2.10}$$

desde $t < t_0$ hasta t_0 obtenemos

$$-\frac{1}{\psi(t_0)} + \frac{1}{\psi(t)} \geq \frac{1}{4} \log \frac{t_0}{t}. \tag{2.11}$$

Cuando $t \longrightarrow 0$ el lado izquierdo está acotado, mientras que el derecho tiende a $+\infty$. Análogamente para $\psi(t_0) > 0$ obtenemos una contradicción. Por lo tanto, θ_0 es admisible.

□

Sin embargo, viendo que el mismo procedimiento para el caso de tres o más dimensiones no funcionaba, Stein se planteó la inadmisibilidad.

2.2. Charles Max Stein

Nacido el 22 de marzo de 1920 en Brooklyn, Nueva York, mostró un talento natural para las matemáticas desde temprana edad. Se consagró como un gigante en su campo, siendo también un ferviente activista social.

Impresionado por Dickson (1929), *Introduction to the Theory of Numbers*, y siendo la Universidad de Chicago su alma máter, inició allí sus estudios.

Obtuvo su licenciatura en matemáticas en 1940. No siguió ningún programa docente fijo por lo que bajo la tutela del Profesor L. M. Graves se introdujo en el análisis. Estudió también el álgebra de Birkhoff y Mac Lane y, se familiarizó con los trabajos de Neyman y Wald, quienes fueron sus grandes referentes.

Sus estudios fueron interrumpidos por la llegada de la Segunda Guerra Mundial, en la que sirvió para las Fuerzas Aéreas. Trabajó la mayor parte del tiempo en el Pentágono para verificar las transmisiones metereológicas, junto con Gilbert Hunt, George Forsythe, Murray Geisler y Kenneth Arrow.

A consecuencia de su desempeño allí, se interesó en graduarse en estadística y obtener el doctorado. Dado el contacto de Geisler y Arrow con Wald, en las clases de Hotelling, quedó atraído por el análisis secuencial de Wald, manteniendo incluso contacto con él por correspondencia.

Stein completó sus estudios el 1947 en la Universidad de Columbia, en contacto con Wald, Hotelling, Wolfowitz y Anderson. Tras marcharse de la Universidad de California, Berkeley, finalmente llegó a Stanford.

Stein fue contratado como profesor asociado en 1953 y, se convirtió en profesor titular tres años más tarde. Siguió viviendo cerca de la universidad durante la mayor parte del resto de su vida.

En 1956, *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, afirmó que la media de una muestra con distribución normal p-variante, como estimador del parámetro θ (la media poblacional) es inadmisible para dimensión $p > 2$, es decir, hay estimadores (necesariamente sesgados) cuyo riesgo es menor que la media muestral $\forall\theta$. En *Estimation with Quadratic Loss* (1961), junto con su ayudante Willard

James, propuso una familia paramétrica de estimadores (de James-Stein) con dicha propiedad de contracción (*Shrinkage*).

Explicando el teorema central del límite en una de sus lecciones, obtuvo una demostración alternativa de este, el Método de Stein. Fue formalizado en *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables* (1972).

2.3. Reacciones

El fenómeno en el que basamos este proyecto fue descrito años más tarde por Efron como "uno de los más sorprendentes teoremas de la estadística matemática de la posguerra".

Sin embargo, impactó a muchos, que como Erich Lehmann, quedaron "aturdidos por la incredulidad". Destacamos la reacción de Dennis L. Lindley, fundador de la Estadística Bayesiana:

> The idea ... is that the mean of a multivariate normal distribution is not best estimated by the sample mean. When I first read of this suggestion several years ago I must admit that I dismissed it as the work of one of these mathematical statisticians who are so entranced by the symbols that they lose touch with reality. It must, I argued, be due to the unbounded loss function, or it could be an improvement, or the sample size was small. But it was none of these things. The estimate proposed by the author is realistic, a great improvement on the sample mean, and makes good practical sense ... We have here one of the most important original statistical ideas of the decade, destined, I feel sure, to influence our thinking and our practice.

También proporcionó otro estimador de la forma

$$\bar{x} + \left(1 - \frac{a_p}{\sum (x_i - \bar{x})^2}\right)(x_i - \bar{x})$$

con a_p una función de p (la dimensión). Obtenía los mismos resultados que James y Stein, pero con un grado menos de libertad. De manera que el mínimo se alcanza en $a_p = p - 3$.

2.4. Encontrando a James

La clase de estimadores presentados en 1961 se denominan estimadores de James-Stein. Notamos la humildad de Stein por citar primero el nombre de su ayudante. Sin embargo, con el transcurso del tiempo la figura de James quedó en el olvido, siendo así que muchos no sabían quién era.

Willard James fue quien ayudó a Stein en la parte computacional para ilustrar los conceptos que llevaba años desarrollando

En los años setenta para esclarecer el efecto de contracción de los estimadores de Stein, Brad Efron y Carl Morris escribieron una serie de artículos. No habían podido identificar quién era W. James dado que había dejado el estado de Fresno (institución listada en el documento original). De manera que tras dieciséis años de búsqueda se dieron por vencidos en poder localizarlo.

No fue hasta 1978, que la American Statistical Asociation (ASA) se puso en contacto con Morris para que participase en una conferencia sobre *Estimación con muchos parámetros: la regla de James-Stein y sus generalizaciones*, y en la que muy a su pesar acabó comentando que los estadísticos no sabían quién era James.

Fue entonces que un hombre de mediana edad sentado en una de las mesas traseras gritó "¡Yo sí lo sé!", a lo que Morris contestó, "¿Y quién es?"
"Soy yo."
Tras una conversación uno a uno a través de la sala, el mundo al fin supo a quién se refería la "W" de "W. James".

Un colega vio el anuncio de la conferencia, y dado que había un James Stein en la facultad le preguntó a este si era a él a quien se referían, a lo que respondió que no era suyo el resultado, sino de Willard James, quien estaba al final del pasillo. Así fue como se enteró y acabó asistiendo a la charla de Morris.

Su colaboración con Stein se dio mientras Willard D. James estudiaba en la Facultad de matemáticas de la Universidad Estatal de California (CSU-LB). Dado que su trabajo de investigación en estadística se había reducido a un verano dedicado a Stein, no había reparado en el impacto que podía tener el artículo a posterior. Confesó sentirse avergonzado de que el estimador se llamara estimador James-Stein, y pidió que se eliminara el "James" para darle el crédito adecuado a Stein.

Doctorado en matemáticas por la Universidad de Illinois el 1957 y asistente de investigación en el departamento de matemáticas de Stanford, conoció a Stein mientras jugaban a juegos de mesa en la hora del almuerzo en el departamento de estadística. En 1959, dado que James buscaba apoyo para su investigación de verano y sabía programar, Stein le propuso que le computara una tabla numérica de los errores medios cuadráticos de su nuevo estimador de contracción, ya que seguramente causaría furor y podrían no creerle.

3. Teoría estadística

En esta sección enumeramos y precisamos los conceptos que empleamos en la memoria.

3.1. ¿Qué tenemos?

Un *modelo estadístico* es una terna $(\mathscr{X}, \mathcal{F}, \mathcal{P})$, donde $(\mathscr{X}, \mathcal{F})$ es un espacio medible y $\mathcal{P}$ es una familia de medidas de probabilidad en $(\mathscr{X}, \mathcal{F})$.

El conjunto al cual pertenecen las observaciones $x_1, ..., x_n$, *espacio muestral*, se denota por $\mathscr{X}$. A $\mathscr{X}$ se le asocia una familia $\mathcal{F}$ de subconjuntos de $\mathscr{X}$ con estructura de σ-álgebra. Habitualmente, $\mathscr{X}$ es un subconjunto de $\mathbb{R}^n$ y $\mathcal{F}$ la σ-álgebra de Borel asociada. Cada observación $x = (x_1, ..., x_n)$ se define como la realización de un objeto aleatorio definido en otro espacio de probabilidad $\tilde{\mathscr{X}}$

$$X = (X_1, ..., X_n) : \tilde{\mathscr{X}} \longrightarrow \mathscr{X}$$

Nos centramos en el estudio de los *modelos estadísticos paramétricos*, que son aquellos cuya familia de probabilidades puede ser descrita como $\{P_\theta : \theta \in \Theta\}$, donde Θ es un subconjunto de $\mathbb{R}^d$, con $d \geq 1$. Es decir, tienen por conjunto de probabilidades a una *familia paramétrica*. Del mismo modo, al conjunto Θ lo denominamos *espacio de parámetros*.

De cada observación solo tenemos referencias del conjunto imagen. De manera que el caso paramétrico, $(\mathscr{X}, \mathcal{F}, P_\theta)$, es el espacio de probabilidad imagen, mediante X, de cierto espacio de probabilidad asociado $\tilde{\mathscr{X}}$ del cual no se hace mención.

3.2. Estimación puntual

El objetivo de la estimación puntual es encontrar a partir de nuestras observaciones un valor aproximado del parámetro θ, o de alguna función de este, $g : \Theta \longrightarrow \mathbb{R}$.

Un *estadístico* es una aplicación medible

$$T : (\mathscr{X}, \mathcal{F}) \longrightarrow (\mathbb{R}^m, \mathscr{B}(\mathbb{R}^m))$$

con $m \geq 1$ entero, y donde $\mathscr{B}$ indica el conjunto de Borelianos.

Un *estimador* de $g(\theta)$ es un estadístico que se usa para estimar $g(\theta)$.

Para valorar la bondad de un estimador podemos estudiar diferentes propiedades. Por ejemplo, un estimador es más $eficiente$ que otro si tiene menor varianza.

El *sesgo* de un estimador integrable T de $g(\theta) \in \mathbb{R}$ es la función

$$\text{Sesgo}_\theta := \mathbb{E}_\theta(T) - g(\theta). \tag{3.1}$$

Consecuentemente, es *insesgado* si

$$\mathbb{E}_\theta(T) = g(\theta), \qquad \forall \theta \in \Theta. \tag{3.2}$$

Se denomina *error cuadrático medio* del estimador T a la función de θ

$$\text{ECM}_\theta(T) := \mathbb{E}_\theta(T - g(\theta))^2 = \text{Var}_\theta(T) + \text{Sesgo}_\theta{}^2(T). \tag{3.3}$$

Para poder definir algunas propiedades interesantes de los estimadores, antes debemos mencionar el concepto de *función de verosimilitud*, que dada una observación $x \in \mathscr{X}$, es por definición

$$\begin{aligned} L(x;\,\cdot\,) : \Theta &\longrightarrow \mathbb{R} \\ \theta &\longmapsto L(x;\theta) = f_\theta(x) \end{aligned}$$

con f la respectiva función de densidad.

Un estadístico T es *suficiente* si la ley del vector X condicionada por el estadístico T no depende de θ. O equivalentemente, si la función de densidad f se puede factorizar en un producto tal que un factor, h, no dependa de θ y el otro factor, que sí depende de θ, dependa de x solo a través de $T(x)$,

$$f(x;\theta) = g(T(x);\theta) \cdot h(x) \qquad c.s.\ \forall x \in \mathscr{X}. \tag{3.4}$$

Por otro lado, un estadístico T es *completo* si $\mathbb{E}_\theta(g(T)) = 0$ para todo $\theta \in \Theta$ implica que $g = 0\ P_\theta^T - c.s.$ para todo $\theta \in \Theta$. Es decir, asegura que las distribuciones correspondientes a diferentes valores de los parámetros sean distintas.

Un modelo es *exponencial* si su función de verosimilitud se puede expresar como

$$L(x;\theta) = exp\{a(\theta)T(X) + b(\theta) + s(x)\}$$

Con b y s funciones conocidas. $T(x)$ se denomina *estadístico canónico o privilegiado.*

Se puede demostrar que un estadístico canónico es suficiente y completo. Véase Corcuera 2019, pp. 27-28.

Un estimador $\hat{\theta}$ de θ es de *máxima verosimilitud* si

$$L(x;\hat{\theta}(x)) = \sup_{\theta\in\Theta} L(x;\theta).$$

Una propiedad útil de estos estimadores es la de *invarianza funcional.* Esta propiedad dice que dado que $\hat{\theta}$ es EMV de θ, entonces dada una biyección g, $g(\hat{\theta})$ es EMV de $g(\theta)$. Esta característica también es válida para el caso multivariante. Es así, que un EMV en el supuesto normal es $(\bar{X}, \frac{1}{n}\sum(X_i - \bar{X})^2)$.

El EMV puede no ser único o incluso no existir, además es encontrado usando un proceso de maximización, lo cual puede conllevar cierta inestabilidad numérica.

3.3. El coste

Para medir la cercanía de un estimador δ al "valor real" de $g(\theta)$, empleamos *funciones de pérdida* $L(\delta, \theta)$, donde, para todo $\theta \in \Theta, L(g(\theta), \theta) = 0$ y $L(\delta, \theta) \geq 0$ en caso contrario. Por lo tanto, no hay pérdida si tomamos la "decisión correcta" $\delta = g(\theta)$ y, es no negativa para cualquier otra decisión. Es común la *función de pérdida cuadrática* definida como

$$L(\delta, \theta) = \|\delta - g(\theta)\|^2 = \sum_{i=1}^{k} (\delta_i - g(\theta_i))^2. \tag{3.5}$$

Se llama *espacio de estimadores de cuadrado integrable* al conjunto

$$\mathscr{D} = \{T \mid \mathbb{E}_\theta(\|T(X)\|^2) < \infty\}, \qquad \forall \theta \in \Theta. \tag{3.6}$$

Y para facilitar la evaluación y tener una idea del *coste* total de un estimador tenemos el concepto de *función de riesgo*, que se define como el valor esperado de la pérdida. Entonces, para un estimador $\delta \in \mathscr{D}$ del parámetro $\theta \in \Theta$ su riesgo definido en $\mathbb{R}^+$ es

$$\mathcal{R}_\delta(\theta) = \mathcal{R}(\delta, \theta) = \mathbb{E}_\theta[L(\delta(X), \theta)]. \tag{3.7}$$

Notamos que en el caso de tomar la función de pérdida cuadrática, su riesgo coincide con el error cuadrático medio.

Finalmente, un estimador es *inadmisible* si existe otro estimador $\delta^* \in \mathscr{D}$ tal que su riesgo sea uniformemente menor, es decir,

$$\mathcal{R}_{\delta^*}(\theta) \leq \mathcal{R}_\delta(\theta), \qquad \forall \theta \in \Theta. \tag{3.8}$$

Si no existe ningún δ^* entonces δ es *admisible*. Cuando la desigualdad es estricta, se dice que δ^* domina a δ.

Por otro lado, δ^* es *minimax* con respecto a la función de riesgo $\mathcal{R}(\delta, \theta)$ asociada si

$$\sup_{\theta \in \Theta} \mathcal{R}(\delta^*, \theta) = \inf_{\delta} \sup_{\theta \in \Theta} \mathcal{R}(\delta, \theta).$$

Es decir, su riesgo máximo es mínimo entre todos los estimadores de θ, o en otras palabras, que δ^* es un estimador que funciona mejor en el peor caso posible permitido en el problema.

4. Stein 1956

Suponemos que estamos interesados en estimar la media θ de una ley Gaussiana y disponemos de una variable aleatoria univariante $X \sim \mathcal{N}(0,1)$. Dada la simetría de las distribuciones normales, la intuición nos dice que observar X es la mejor manera de estimar θ. Para evaluar la calidad de nuestro estimador intuitivo nos basamos principalmente en dos criterios, una función de pérdida cuadrática y su riesgo asociado, en este caso el error cuadrático medio.
Inicialmente, la dominancia sobre X por parte de los estimadores James-Stein (JS) fue sorprendente, ya que el propio X es:

- el mejor estimador insesgado de θ.
- el mejor estimador invariante por translaciones de θ,
- el estimador de máxima verosimilitud (EMV) de θ,
- un estimador minimax de θ, y
- un estimador admisible de θ cuando $p = 1$ o 2.

Sin embargo, Stein probó que X era inadmisible para $p \geq 3$ (1956) y proporcionó estimadores que lo dominaban (1961). Estos eran de la forma $X + g(X)$, con g una cierta función, también minimax, tales que su riesgo es $p\sigma^2$ menos un término proporcional a $\frac{1}{\|X\|^2}$, de manera que el riesgo es igual al de X cuando su norma tiende a infinito y menor para todo θ finito.

Así pues, el estimador de James-Stein es un estimador de la media, θ, de una distribución normal que, al precio de incorporar sesgo, domina el estimador habitual bajo un error cuadrático medio.

4.1. El Lema de Stein

4.1.1. Caso univariante

Lema 4.1.1. *Sea X una variable aleatoria real con ley $\mathcal{N}(0,1)$ y $g : \mathbb{R} \longrightarrow \mathbb{R}$ una función derivable. Entonces, si $\mathbb{E}|g'(X)| < \infty$,*

$$\mathbb{E}[Xg(X)] = \mathbb{E}[g'(X)].$$

Demostración. Integrando por partes:

$$u = g(x), \qquad du = g'(x)dx,$$
$$v = -e^{-\frac{1}{2}x^2}, \qquad dv = xe^{-\frac{1}{2}x^2}.$$

Tenemos

$$\begin{aligned}
\mathbb{E}[Xg(X)] &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} x g(x) dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{1}{2}x^2} (x-\mu) g(x) dx \\
&= \frac{1}{\sqrt{2\pi}} \left(\int_{\mathbb{R}} e^{-\frac{1}{2}x^2} g'(x) dx - \left[g(x) e^{-\frac{1}{2}x^2} \right]_{\mathbb{R}} \right) \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{1}{2}x^2} g'(x) dx = \mathbb{E}(g'(X)).
\end{aligned}$$

□

Corolario. *Dada una variable aleatoria* $Y \sim \mathcal{N}(\mu, \sigma^2)$, $h(Y) = g\left(\frac{Y-\mu}{\sigma}\right)$ *y bajo las hipótesis del lema anterior,*

$$\mathbb{E}[(Y-\mu)h(Y)] = \sigma^2 \mathbb{E}[h'(Y)]. \tag{4.1}$$

Demostración.

$$\begin{aligned}
\mathbb{E}[h'(Y)] &= \frac{1}{\sigma} \mathbb{E}\left[g'\left(\frac{Y-\mu}{\sigma} \right) \right] \\
&= \frac{1}{\sigma} \mathbb{E}[g'(X)] = \frac{1}{\sigma} \mathbb{E}[Xg(X)] \\
&= \frac{1}{\sigma} \mathbb{E}\left[\frac{Y-\mu}{\sigma} g\left(\frac{Y-\mu}{\sigma} \right) \right] \\
&= \mathbb{E}\left[\frac{Y-\mu}{\sigma^2} h(Y) \right].
\end{aligned}$$

□

4.1.2. Caso multivariante

Para garantizar la integración por partes introducimos el siguiente concepto, menos estricto que el de diferenciabilidad, y que esencialmente hace referencia a funciones *débilmente diferenciables*.

Definición 4.1.1. Denominamos a una función $g : \mathbb{R}^p \longrightarrow \mathbb{R}$ como *casi diferenciable* si existe una función $\nabla g : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ tal que para todo $a \in \mathbb{R}^p$,

$$g(x+a) - g(x) = \int_0^1 a \cdot \nabla g(x+ta) dt.$$

Esencialmente, el operador ∇g es el vector de derivadas parciales con $\nabla_i = \frac{\partial}{\partial_i}$. De modo que si la función es dos veces continuamente diferenciable definimos el *Laplaciano* como $\Delta g = \nabla^2 g = \sum_{i=1}^{p} \nabla_i^2 g_i$.

Si en esta sección $g : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ es casi diferenciable, escribiremos

$$\operatorname{div} g = \nabla \cdot g = \sum_{i=1}^{p} \nabla_i g_i$$

.

Lema 4.1.2. *Dada $X \sim \mathcal{N}_p(\theta, I_p)$ una variable aleatoria p-variante, y $g : \mathbb{R}^p \longrightarrow \mathbb{R}$ una función casi diferenciable tal que $\mathbb{E}|\nabla g(X)| < \infty$. Entonces,*

$$\mathbb{E}[(X-\theta)'g(X)] = \mathbb{E}[\nabla' g(X)]. \tag{4.2}$$

Demostración. Para $i = 1, ..., p$, la notación $X = (X_i, X_{-i})$ representa el vector aleatorio en el que separamos la componente *i-ésima* del resto de componentes. Usando la independencia de X_i y X_{-i} y el lema anterior, obtenemos

$$\mathbb{E}[(X_i-\theta_i)'g(X)|X_{-i}] = \mathbb{E}[(\nabla g)_i'(X)|X_{-i}].$$

Y tomando esperanzas a ambos lados, tenemos

$$\mathbb{E}[(X_i-\theta_i)'g(X)] = \mathbb{E}[(\nabla g)_i'(X)], \qquad \forall i = 1, ..., p.$$

□

4.2. Construcción del estimador

Diferenciamos distintos casos en función de la información que disponemos de la varianza de X.

4.2.1. Caso 1. Varianza conocida y constante

Dada $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$, suponemos para simplificar cálculos que el vector de medias θ es igual a 0 y, la matriz de varianzas y covarianzas es la identidad.

Teorema 4.2.1. *Dada una función diferenciable $g : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ tal que*

$$\mathbb{E}\left[\sum_{i=1}^{p} |\nabla_i g_i(X)|\right] < \infty.$$

Entonces,

$$\mathbb{E}[\|X + g(X)\|^2] = p + \mathbb{E}[2\nabla \cdot g(X) + \|g(X)\|^2]. \tag{4.3}$$

Demostración. Usando el lema anterior sobre g_i, tenemos

$$\begin{aligned}\mathbb{E}[\|X_i + g_i(X)\|^2] &= \mathbb{E}[X_i^2 + 2X_i' g_i(X) + g_i^2(X)] \\ &= 1 + 2\mathbb{E}[\nabla' g_i(X)] + \mathbb{E}[(g_i^2(X))].\end{aligned}$$

Sumamos por cada $i \in \{1, ..., p\}$ y obtenemos el resultado. □

Tratamos de encontrar un estimador de la forma $X + g(X)$, de manera que su riesgo quede descompuesto entre el error del estimador usual y un término negativo que dependa de la función g, haciendo que tengamos una pérdida cuadrática inferior. Nos centramos en las funciones $g : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ del tipo

$$g = \nabla \log f = \frac{\nabla f}{f}$$

Con f bien definida. Esto nos lleva a una nueva versión del teorema anterior.

Teorema 4.2.2. *Dada una función casi diferenciable $f : \mathbb{R}^p \longrightarrow \mathbb{R}_0^+$ con ∇f casi diferenciable, y tal que*

$$\mathbb{E}\left[\frac{1}{f(X)}\sum_{i=1}^{p}|\nabla_i^2 f(X)|\right] < \infty, \qquad \mathbb{E}[\|\nabla \log f(X)\|^2] < \infty.$$

Entonces,

$$\begin{aligned}\mathbb{E}[\|X + \nabla \log f\|^2] &= p + \mathbb{E}\left[2\frac{\nabla^2 f(X)}{f(X)} - \frac{\|\nabla f(X)\|^2}{f^2(X)}\right] \\ &= p + 4\mathbb{E}\left[\frac{\nabla^2\sqrt{f(X)}}{\sqrt{f(X)}}\right]. \end{aligned} \tag{4.4}$$

Demostración. Sea $g : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ definida como $g = \nabla log f = \dfrac{\nabla f}{f}$. Entonces,

$$\nabla \cdot g = \nabla \cdot (\nabla \log f) = \frac{\nabla^2 f}{f} - \frac{\|\nabla f\|^2}{f^2},$$

y a partir del teorema anterior,

$$\begin{aligned}\mathbb{E}[\|X + \nabla \log f\|^2] &= p + \mathbb{E}\left[\frac{\|\nabla f\|^2}{f^2} + 2\left(\frac{\nabla^2 f}{f} - \frac{\|\nabla f\|^2}{f^2}\right)\right] \\ &= p + \mathbb{E}\left[2\frac{\nabla^2 f}{f} - \frac{\|\nabla f\|^2}{f^2}\right].\end{aligned}$$

Finalmente, dado que

$$\nabla^2(\sqrt{f}) = \nabla \cdot (\nabla\sqrt{f}) = \nabla \cdot \frac{\nabla f}{2\sqrt{f}} = \frac{\nabla^2 f}{2\sqrt{f}} - \frac{\|\nabla f\|^2}{4f^{\frac{3}{2}}},$$

Substituimos $\|\nabla f\|^2$ por $2f\nabla^2 f - 4f^{\frac{3}{2}}\nabla^2\sqrt{f}$ y obtenemos

$$\begin{aligned}\mathbb{E}[||X + \nabla \log f||^2] &= p + \mathbb{E}\left[2\frac{\nabla^2 f}{f} - \frac{\|\nabla f\|^2}{f^2}\right] \\ &= p + \mathbb{E}\left[2\frac{\nabla^2 f}{f} - \frac{2f(\nabla^2 f) - 4f^{\frac{3}{2}}(\nabla^2\sqrt{f})}{f^2}\right] \\ &= p + 4\mathbb{E}\left[\frac{\nabla^2\sqrt{f}}{\sqrt{f}}\right].\end{aligned}$$

□

Así pues, debemos encontrar una función f que cumpla las hipótesis del teorema y que $\nabla^2\left(\sqrt{f(x)}\right) \leq 0$, de manera que tendremos un estimador de la forma $X + \nabla \log f(X)$ que dominará al estimador usual del vector de medias. De hecho, en este caso,

$$\mathbb{E}[\|X + \nabla \log f(X) - \theta\|^2] \leq p = \mathbb{E}[\|X - \theta\|^2].$$

El estimador de James-Stein lo obtenemos precisamente tomando una función del tipo $f(x) = \left(\dfrac{1}{\|X\|^2}\right)^b$. De manera que

$$\nabla f(X) = -b\left(\|X\|^2\right)^{-(b+1)} 2X,$$

y

$$\nabla \log f(X) = \frac{\nabla f(X)}{f(X)} = \frac{-2b}{\|X\|^2} X.$$

Y dado que $\nabla^2 \left(\sqrt{f(X)}\right) = -\frac{b(p-2-b)}{\|X\|^{b+2}}$,

$$\delta_b^{JS}(X) = \left(1 - \frac{2b}{\|X\|^2}\right) X. \tag{4.5}$$

Este estimador domina a X para $0 \leq b \leq (p-2)$ (si $p > 2$). Su riesgo es $p - 4\mathbb{E}\left[\frac{b(p-2-b)}{\|X\|^2}\right]$, siendo este mínimo en $b = \frac{p-2}{2}$.

Resiguiendo estos cálculos y substituyendo X por $\frac{X-\theta}{\sigma}$ definimos el estimador de James-Stein para cualquier vector de medias y varianza un múltiplo conocido de la identidad de la siguiente manera.

Definición 4.2.1. Dada $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$, con σ^2 conocida, el *estimador de James-Stein* es

$$\delta^{JS}(X) = \left(1 - \frac{(p-2)\sigma^2}{\|X\|^2}\right) X. \tag{4.6}$$

Con riesgo igual a

$$\sigma^2 \left(p - \sigma^2 (p-2)^2 \mathbb{E}\left[\frac{1}{\|X\|^2}\right]\right). \tag{4.7}$$

Así pues, la Paradoja de Stein, presentada en 1956, la formalizamos de la siguiente manera.

Teorema 4.2.3. *El estimador usual de la media de una variable aleatoria normal $p-$variante, $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$, con σ conocida, es inadmisible para $p \geq 3$.*

Demostración. Veremos que bajo pérdida cuadrática, el EMV usual, X, de θ tiene un riesgo superior al del estimador de James-Stein, siendo entonces inadmisible.
Primeramente, sabemos que

$$\mathrm{ECM}\,(X) = \mathbb{E}\left[\sum_{i=1}^{p} (X_i - \theta_i)^2\right] = p\sigma^2.$$

Luego, dado que $(p-2)^2 > 0$ y que $\|X\|^2 \sim \mathcal{X}^2{}_p(\|\theta\|^2)$, se distribuye igual que una variable ji cuadrado centrada tomando una distribución de Poisson para los grados de libertad. Es decir, $\|X\|^2 \sim \mathcal{X}^2{}_{p+2K}$ donde $K \sim Poisson(\|\theta\|^2/2)$. Tenemos que,

$$\mathbb{E}\left[\frac{1}{\|X\|^2}\right] = \mathbb{E}\left[\frac{1}{\mathcal{X}^2_{p+2K}}\right] = \mathbb{E}\left[\frac{1}{p-2+2K}\right].$$

De este modo,

$$\sigma^2 \left(p - \sigma^2 (p-2)^2 \mathbb{E}\left[\frac{1}{\|X\|^2}\right]\right) < p\sigma^2. \tag{4.8}$$

Haciendo así que X no sea admisible para $p \geq 3$. □

Observación 4.2.1. Cuanto más cercana a 0 sea $\|X\|^2$ menor riesgo tendremos y, por lo tanto una mayor mejora y, conforme la norma aumente más cercano será δ^{JS} al valor real de X. En consecuencia, este estimador nos será útil en los casos en que $\|\theta\|$ sea pequeña, por lo que nos resultará favorable primero centralizar los datos. Ampliamos esta observación en la Sección 5.

Observación 4.2.2. Este resultado no implica que el estimador de James-Stein sea admisible, de hecho no lo es.

4.2.2. Caso 2. Varianza constante y desconocida

Para facilitar los cálculos consideramos el vector de medias nulo, es decir, $X \sim \mathcal{N}_p(0, \sigma^2 I_p)$, con σ^2 desconocida. Tomamos $Y = (Y_1, ..., Y_{n+1})$, con $Y_j \sim \mathcal{N}(0, \sigma^2)$ componentes de X_i (ya que las p componentes tienen la misma distribución). De este modo, $\bar{Y} = \dfrac{\sum_{j=1}^{n+1} Y_j}{n+1}$.

Dado que desconocemos el valor de σ, usamos el Teorema de Fisher (Corcuera, 2019) sobre Y para estimar su valor a partir de la varianza muestral corregida S^2. Así pues, $\upsilon \equiv nS^2 \sim \sigma^2 \mathcal{X}_n^2$. Con $\bar{Y}$ y S^2 independientes.

Siguiendo de base los cálculos anteriores, ahora observamos el riesgo de los estimadores de la forma $\left(1 - \dfrac{c(p-2)\upsilon}{\|X\|^2}\right) X$, con c una constante a determinar.

Dado que $\mathcal{X}_n^2 = \displaystyle\sum_{i=1}^{n} Z_i^2$, con $Z_i \sim \mathcal{N}(0,1)$, y Z_i son independientes,

$$\begin{aligned}\mathbb{E}[\upsilon^2] = \sigma^2 \mathbb{E}\left[(\mathcal{X}_n^2)^2\right] &= \mathbb{E}\left[\sum_{i}^{n} Z_i^4 + 2\sum_{i<j}^{n} Z_i^2 Z_j^2\right] \\ &= \sigma^2\left(3n + 2\frac{n(n-1)}{2}\right) = \sigma^2(n^2+2n).\end{aligned}$$

De este modo y usando el lema de Stein con $g(X) = \dfrac{X}{\|X\|^2}$, tenemos

$$\mathbb{E}\left[\left\|X - \frac{c(p-2)\upsilon}{\|X\|^2} X\right\|^2\right]$$

$$\begin{aligned}&= \mathbb{E}\left[\|X\|^2 - X' 2c(p-2)\upsilon \frac{X}{\|X\|^2} + c^2(p-2)^2\upsilon^2 \frac{1}{\|X\|^2}\right] \\ &= \sigma^2 p - 2c(p-2)\mathbb{E}[\upsilon]\mathbb{E}\left[\frac{X'X}{\|X\|^2}\right] + c^2(p-2)^2\mathbb{E}[\upsilon^2]\mathbb{E}\left[\frac{1}{\|X\|^2}\right] \\ &= \sigma^2\left(p - 2c(p-2)n\mathbb{E}\left[\frac{X'X}{\|X\|^2}\right] + \sigma^2 c^2(p-2)^2 n(n+2)\mathbb{E}\left[\frac{1}{\|X\|^2}\right]\right). \\ &= \sigma^2\left(p - n\sigma^2(p-2)^2(2c - c^2(n+2))\mathbb{E}\left[\frac{1}{\|X\|^2}\right]\right).\end{aligned}$$

Siendo el riesgo mínimo en $c = \dfrac{1}{n+2}$.

Definición 4.2.2. Dada $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$, con varianza desconocida estimada por $\upsilon \sim \sigma^2 \mathcal{X}_n^2$, el estimador de James-Stein es

$$\delta^{JS}(X) = \left(1 - \frac{(p-2)\upsilon}{(n+2)\|X\|^2}\right) X. \tag{4.9}$$

También podríamos haber tomado $c = \dfrac{1}{n}$, menos óptimo, que nos daría el estimador

$$\left(1 - \frac{(p-2)\upsilon}{n\|X\|^2}\right) X. \tag{4.10}$$

Donde como $\mathbb{E}\left[\dfrac{\upsilon}{n}\right] = \sigma^2$, simplemente se trata de 4.6.

4.2.3. Caso 3. Varianza conocida, generalización

Análogamente tratamos el caso $X \sim \mathcal{N}_p(\theta, \Sigma)$, $\Sigma = \operatorname{diag}(\sigma_1^2, ..., \sigma_p^2)$ conocida. Hacemos el mismo tipo de transformación que en el caso anterior, $\upsilon_i = \sigma_i^2 \mathcal{X}_{n_i}^2$, para todo $i \in \{1, ..., p\}$. De este modo obtenemos el siguiente estimador.

Definición 4.2.3. Dada $X \sim \mathcal{N}_p(\theta, \Sigma)$, con matriz de covarianza $\Sigma = \operatorname{diag}(\sigma_1^2, ..., \sigma_p^2)$ conocida, el estimador de James-Stein se define como

$$\delta_i^{JS}(X) = \left(1 - \frac{(p-2)\upsilon_i}{n_i\|X\|^2}\right) X_i, \qquad \forall i = 1, ..., p. \tag{4.11}$$

4.2.4. Caso 4. Varianza desconocida, generalización

Suponemos $X \sim \mathcal{N}_p(\theta, \Sigma)$ con Σ una matriz desconocida y semidefinida positiva. Análogamente, tomamos una muestra de tamaño n, $Y_1, ..., Y_n, (Y_i \sim \mathcal{N}_p(0, \Sigma))$ y definimos $S = \sum Y_i Y_i'$. En este caso consideramos la función de pérdida

$$\mathbb{E}\left[(\hat{\theta} - \theta)'\Sigma^{-1}(\hat{\theta} - \theta)\right].$$

Suponiendo que S es invertible obtenemos el siguiente estimador.

Definición 4.2.4. Dada $X \sim \mathcal{N}_p(\theta, \Sigma)$, con matriz de covarianza $\Sigma = \operatorname{diag}(\sigma_1^2, ..., \sigma_p^2)$ conocida, el estimador de James-Stein se define como

$$\delta_i^{JS}(X) = \left(1 - \frac{p-2}{(n-p+3)X'S^{-1}X}\right) X. \tag{4.12}$$

Con riesgo

$$p - \frac{n-p+1}{n-p+3}(p-2)^2\mathbb{E}\left[\frac{1}{\|X\|^2}\right].$$

4.3. Mejoras

A raíz de las observaciones 4.2.1 y 4.2.2, vemos que cuando $\|X\|^2$ es pequeño, el término que multiplica a X en 4.6 puede ser negativo, lo que podría causar que el estimador no encogiera. Una solución simple es forzar el coeficiente a ser positivo, restringiéndolo a su parte positiva.

Proposición 4.3.1. *Si $\theta = \mathbb{E}[X]$ es estimado por $h(X)X$, donde $P(h(X) < 0) > 0$. Bajo la hipótesis adicional de que $\mathbb{E}[X_i|h(x) < 0]$ tiene el mismo signo que θ_i para todo $i \in \{1, ..., p\}$, donde p es la dimensión de X, y definiendo $h_+(X)X = max(0, h(X))$, tenemos*

$$\text{ECM}\,(h_+(X)X) \leq \text{ECM}\,(h(X)X). \tag{4.13}$$

Demostración. Observamos la contribución de $h(x)$ en las componentes del ECM ,

$$\mathbb{E}[(h(x)X_i - \theta_i)^2] = \mathbb{E}[h^2(X)X_i^2] - 2\theta_i\mathbb{E}[h(x)X_i] + \theta_i^2. \tag{4.14}$$

Ignorando el último término por ser positivo, condicionamos respecto al signo de $h(X)$,

$$\begin{aligned}
&\mathbb{E}[h^2(X)X_i^2] - 2\theta_i\mathbb{E}[h(x)X_i] \\
&= (\mathbb{E}[h^2(X)X_i^2|h(X) \geq 0] - 2\theta_i\mathbb{E}[h(x)X_i|h(X) \geq 0])P(h(X) \geq 0) \\
&+ (\mathbb{E}[h^2(X)X_i^2|h(X) < 0] - 2\theta_i\mathbb{E}[h(x)X_i|h(X) < 0])P(h(X) < 0).
\end{aligned}$$

Como proponemos reemplazar $h(X)$ por $h_+(X)$, la única diferencia se da cuando $h(X) < 0$, por lo que solo debemos fijarnos en la segunda parte de la fórmula. Entonces tenemos

$$\mathbb{E}[h^2(X)X_i^2|h(X) < 0] \geq 0,$$

y teniendo en cuenta la hipótesis adicional del enunciado, obtenemos

$$-2\theta_i\mathbb{E}[h(x)X_i|h(X) < 0]) > 0.$$

Así que 4.14 sería más pequeño si reemplazáramos $h(X)$ por 0 si este fuera negativo, de ahí la dominancia del estimador $h_+(X)X$. □

En nuestro caso, $h(X) = \left(1 - \dfrac{b}{\|X\|^2}\right)$ y la hipótesis adicional se cumple, ya que

$$\begin{aligned}
\mathbb{E}[X_i|h(X) < 0] &= \mathbb{E}[X_i|\|X\|^2 < b] \\
&= \mathbb{E}\left[\mathbb{E}[X_i|X_i^2]\middle|\sum_{i\neq k} X_i^2 < b - X_i^2\right] \\
&= \mathbb{E}\left[|X_i|P(X_i > 0|X_i^2) - |X_i|P(X_i < 0|X_i^2)\middle|\sum_{i\neq k} X_i^2 < b - X_i^2\right]
\end{aligned} \tag{4.15}$$

tiene el mismo signo que θ_i porque, si $\theta_i > 0$, entonces $P(X_i > 0|X_i^2) > P(X_i < 0|X_i^2)$ y por lo tanto 4.15 es también positivo. Análogamente, si $\theta_i < 0$ tenemos que 4.15 es negativo.

Esto nos lleva al siguiente estimador, normalmente llamado *estimador de la parte positiva de James-Stein*,

$$\delta_+^{JS} = \left(1 - \frac{p-2}{\|X\|^2}\right)_+ X. \tag{4.16}$$

Asimismo, la mejora de cuando X está cerca del origen aún podemos perfeccionarla para $n \geq 1$ observaciones centrando los datos, de manera que perdemos un grado de libertad y el estimador se convierte en

$$\left(1 - \frac{(p-3)\sigma^2}{n\|X - \bar{X}\|^2}\right)(X - \bar{X}) + \bar{X}$$

En vez de contraer hacia el origen, ahora encogemos hacia $\bar{X}$. A este estimador le llamamos *estimador de James-Stein centrado*. Cabe remarcar que no deja de ser un estimador de Bayes empírico de James-Stein. Esto nos aporta una mejora cuando la media es cercana a un múltiple de $(1, ..., 1)'$, o cuando todas las observaciones de nuestra muestra tienen medias similares. En todo caso, dependiendo de la forma de la media, puede darnos peores resultados.

Visualizamos todos estos resultados en la siguiente sección.

5. Simulaciones

Las simulaciones las hemos hecho con `R` (el código está en el anexo) y exploran las diferentes situaciones que estamos estudiando.

Remarcamos primeramente que el estimador δ^{JS} no contrae cada componente hacia la media, por lo que podemos encontrarnos componentes que tengan un error cuadrático mayor, sin embargo, al tomar su riesgo esa diferencia se ve compensada por las mejoras de las demás componentes. Para mostrarlo observamos la Figura 1, en la que $\theta = (1, ..., 1)$ de dimensión $p = 20$.

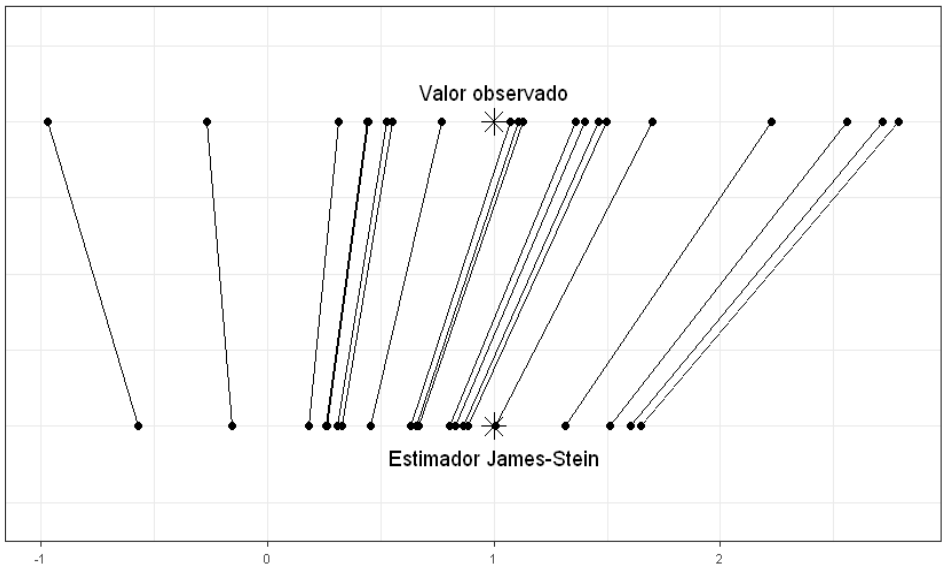


Figura 1: Contracción hacia el origen

Para cada gráfico generamos una muestra de tamaño 50000 para estimar el ECM , y salvo que indiquemos lo contrario, establecemos $p = 20$. Iniciamos considerando el caso multinormal con la identidad como matriz de covarianza.

En la Figura 2 vemos que el valor observado X tiene un ECM igual a la dimensión mientras que δ^{JS} cancela completamente el aumento lineal del ECM con un valor constante de 2,5 aproximadamente, e incluso algo mejor al tomar su versión positiva. La mejora que se muestra es bastante natural, ya que δ^{JS} puede verse como una contracción hacia el origen.

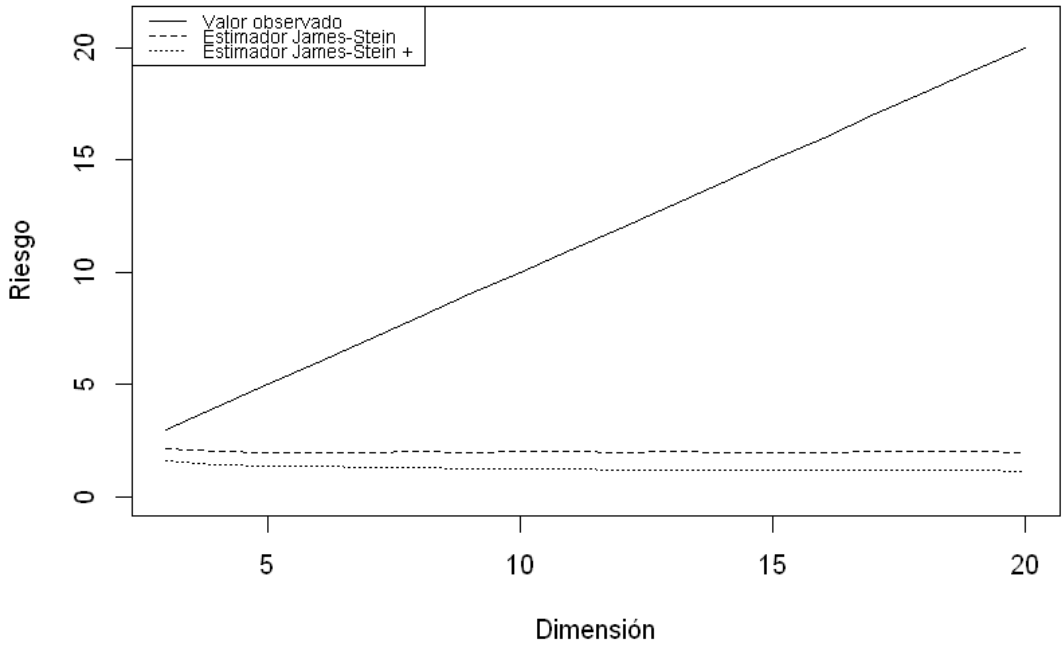


Figura 2: ECM según la dimensión, con $\theta = (0, ..., 0)$

Es más paradójico cuando la media no coincide con el origen, ya que como la ley Gaussiana es simétrica no parece natural contraer hacia 0, pero observamos que cuando la media es $(1, ..., 1)$ la mejora es aún sustancial. Vemos también que δ_+^{JS} y δ^{JS} casi coinciden, ya que las diferencias solo se dan cuando la norma de X es realmente pequeña, que es con menos frecuencia en este caso.

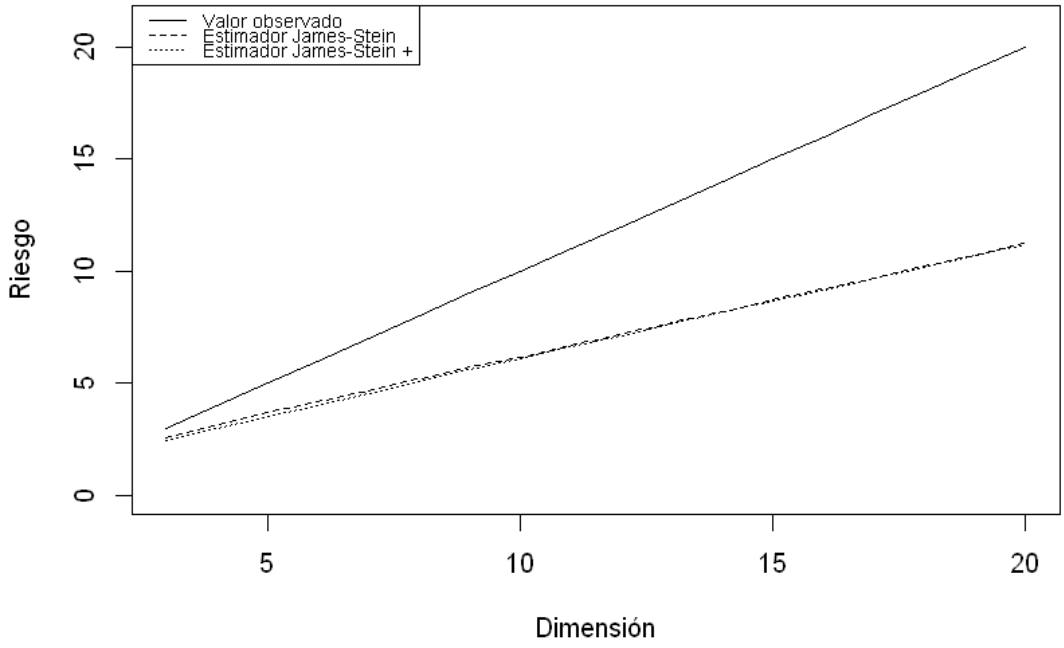


Figura 3: ECM según la dimensión, con $\theta = (1, ..., 1)$

Si tomamos $n \geq 1$ observaciones de cada muestra, δ^{JS} pasa a obtenerse a partir de

$$\left(1 - \frac{(p-2)\sigma^2}{n\|\bar{X}\|^2}\right)\bar{X}, \tag{5.1}$$

donde $\bar{X}$ es la matriz de medias de dimensión $N \times p$ de cada n. De este modo fijando $n = 10$ y suponiendo X Gaussiana obtenemos el siguiente gráfico, en el que de manera natural el riesgo del EMV es $\frac{p\sigma^2}{n}$,

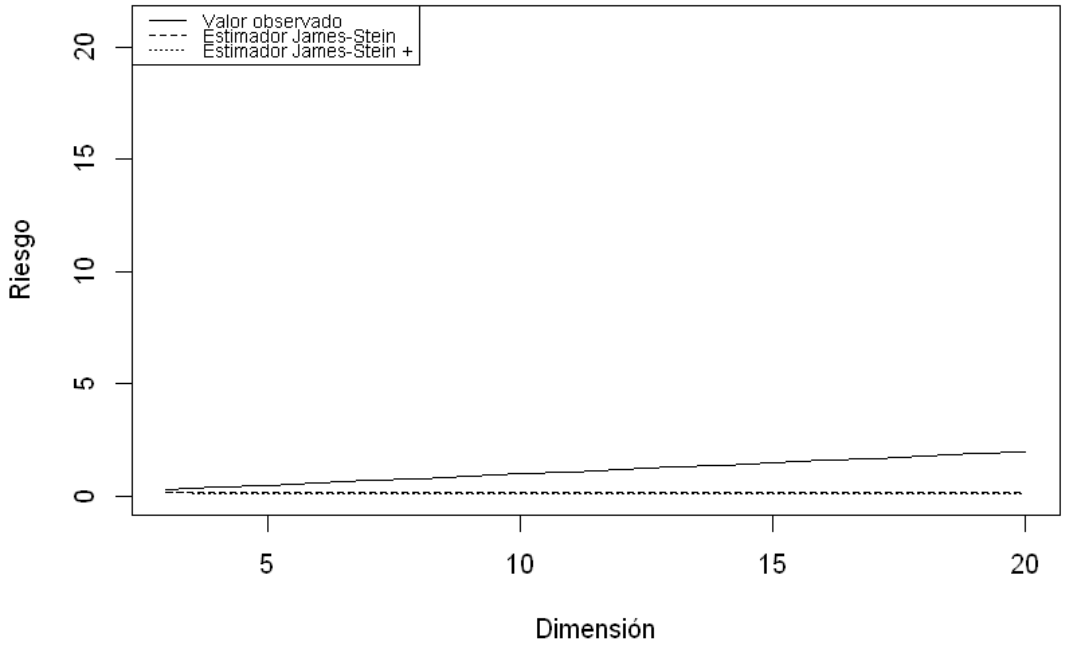


Figura 4: ECM según la dimensión, con $\theta = (0, ..., 0)$ y $n = 10$

Si además, suponemos que $\theta \sim \mathcal{N}(\nu, \tau^2)$ comparando los casos $(0, 1)$ y $(0, 5^2)$, observamos que cuanta mayor varianza tenga el vector de medias, más se aproximan nuestros estimadores al EMV , y por lo tanto, menor mejora obtenemos.

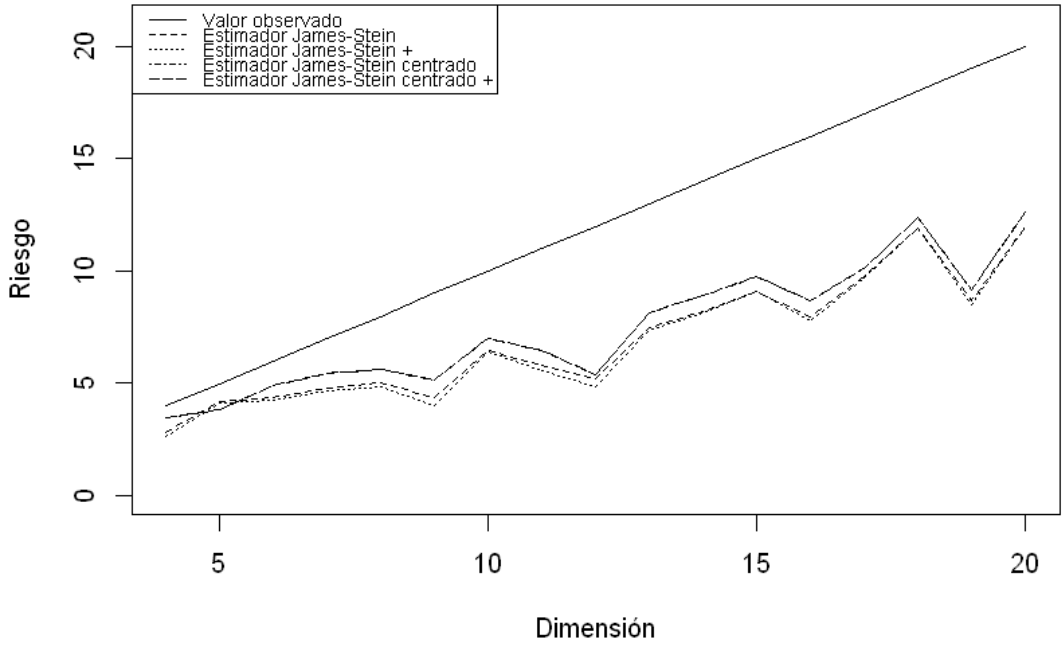


Figura 5: ECM según la dimensión, con $\theta \sim \mathcal{N}(0,1)$ y $n = 1$

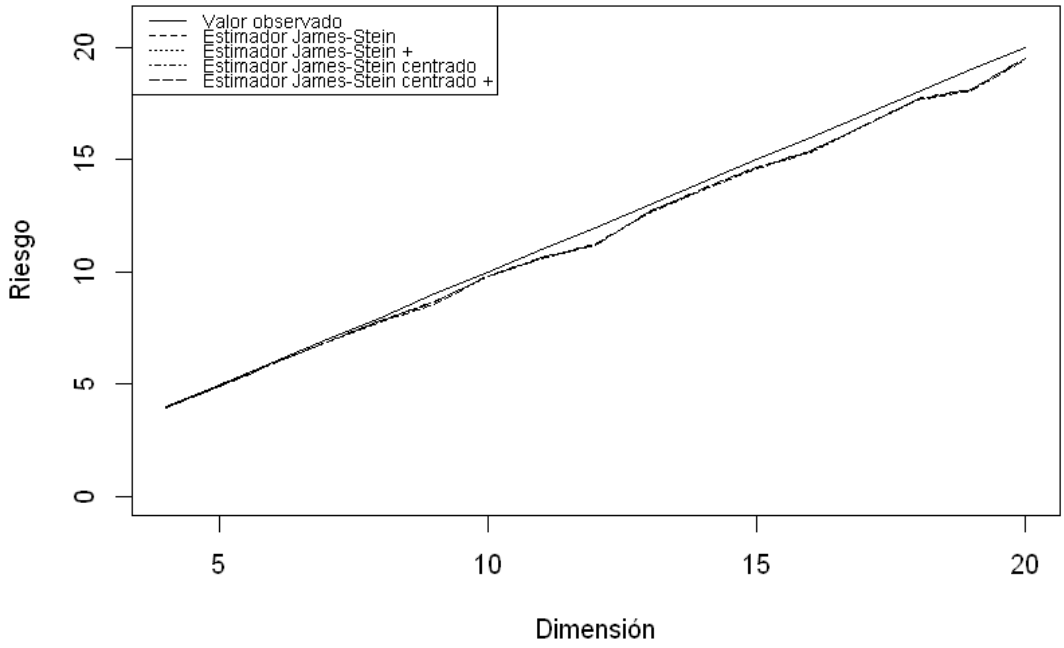


Figura 6: ECM según la dimensión, con $\theta \sim \mathcal{N}(0,25)$ y $n = 1$

Fijamos ahora la dimensión a 20 y consideramos $\sigma^2 = 1$ para centrarnos en la influencia de la norma y en la forma del vector de medias.

Vemos que los ECM se acercan al de X conforme la norma de la media aumenta. También observamos que la forma de la media solo influye en los estimadores centrados cuando: - $\theta = m \cdot (1, ..., 1)'$. La versión centrada da el mismo resultado que si $\theta = 0$ (lo cual era de esperarse), y la versión centrada positiva es aún mejor.

- $\theta = m \cdot (1, 2, ..., p)'$. La versión centrada nos da un buen margen.

También observamos ligeras mejoras con $\theta \sim \mathcal{N}(1, 1)$ o $\theta \sim \text{Pois}(1)$.

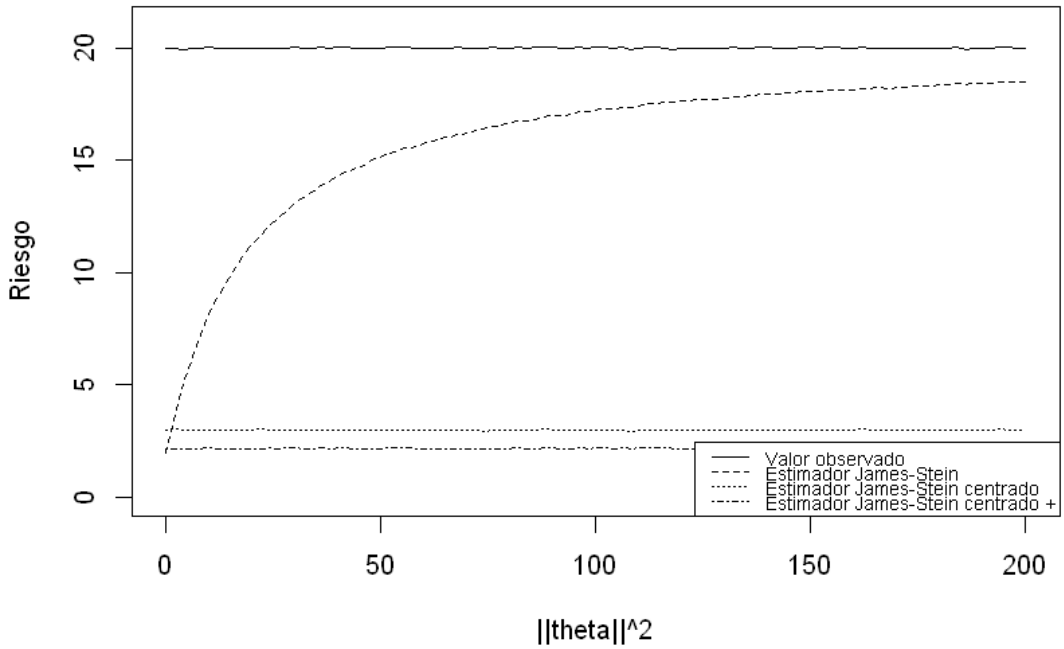


Figura 7: ECM cuando la norma de la media aumenta, con $\theta = m \cdot (1, ..., 1)'$

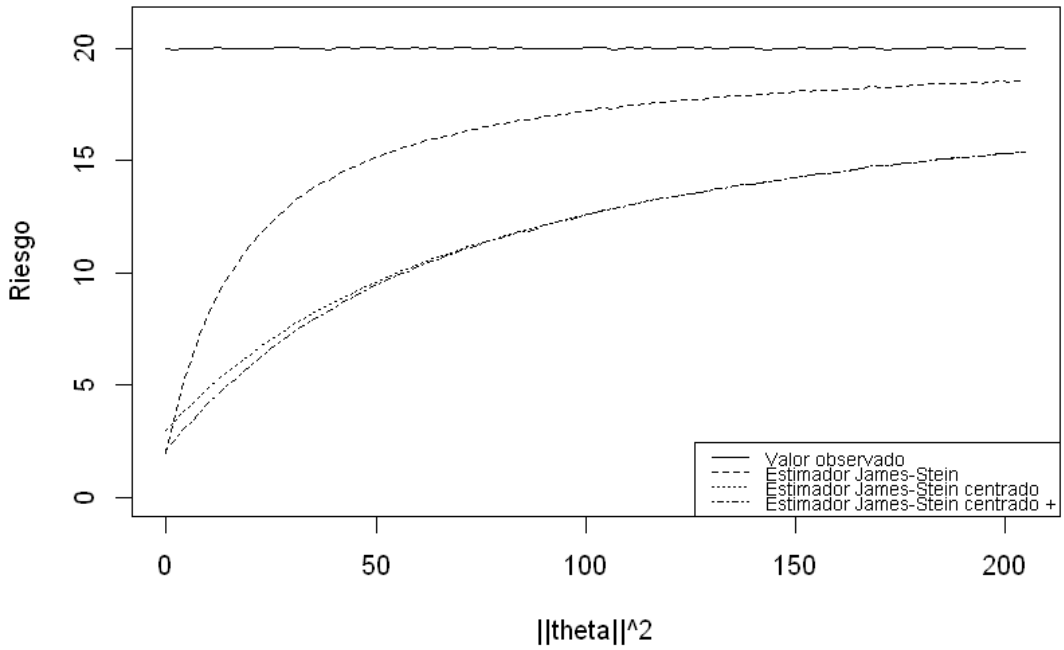


Figura 8: ECM cuando la norma de la media aumenta, con $\theta = m \cdot (1, 2, ..., 20)'$

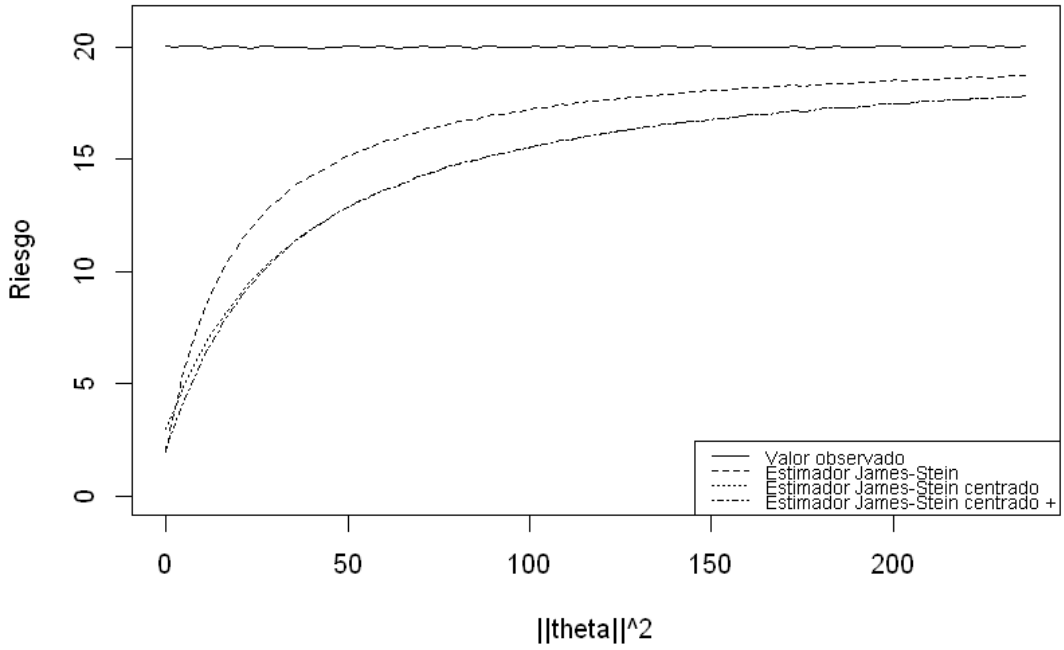


Figura 9: ECM cuando la norma de la media aumenta, con $\theta \sim \mathcal{N}(1,1)$ y $n = 1$

Sin embargo, centrar no siempre es una buena solución para medias distintas de 0, por ejemplo con $\theta \sim \mathcal{N}(0,1)$, $\theta \sim \mathcal{N}(0,25)$, $\theta \sim \mathcal{N}(1,25)$ o $\theta = m \cdot (0,...,0,1)'$. Para apreciar mejor esta diferencia bajamos la dimensión a 5.

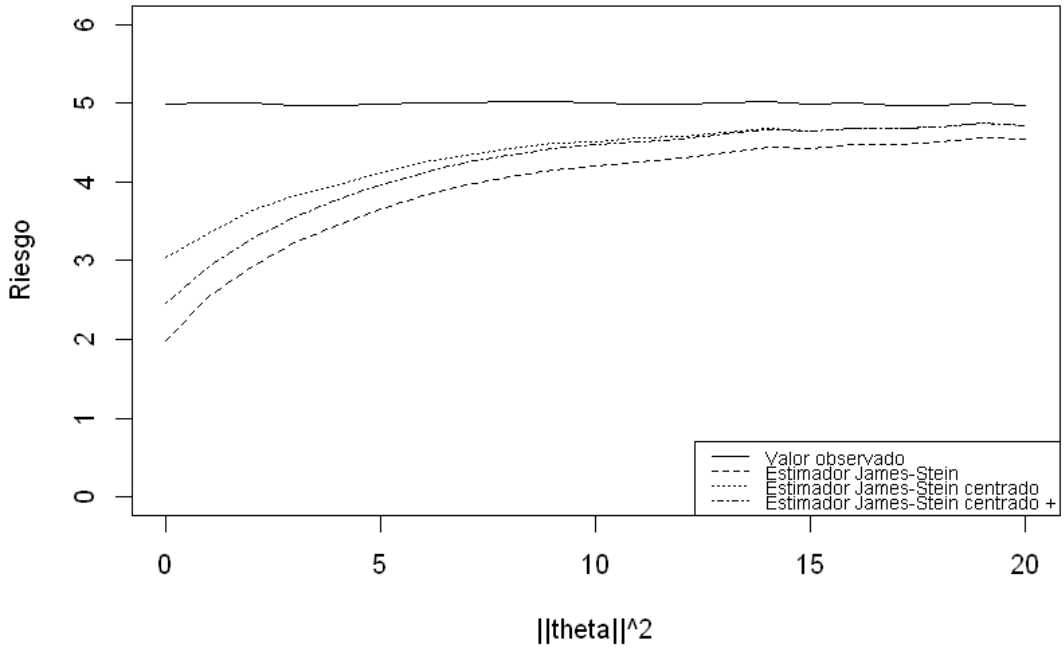


Figura 10: ECM cuando la norma de la media aumenta, con $\theta = m \cdot (0,...,0,1)'$)

En el caso donde la matriz de covarianza es un múltiplo desconocido de la identidad, $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$, consideramos los estimadores 4.9 y 4.10 de la Sección 4. Tomamos un valor pequeño de n, $n = 10$, para mostrar que tomar el denominador igual a $n+2$ es mejor que n. Si n aumenta, la diferencia se vuelve insignificante. Entonces, tomamos $p = 20$, y $\theta = (1,...,1)'$ y con σ^2 variable. La v encontrada en los estimadores estudiados sería

normalmente obtenida a partir de una muestra, pero para simplificar las simulaciones, la generamos directamente como $\sigma^2 \mathcal{X}_n^2$ variable.

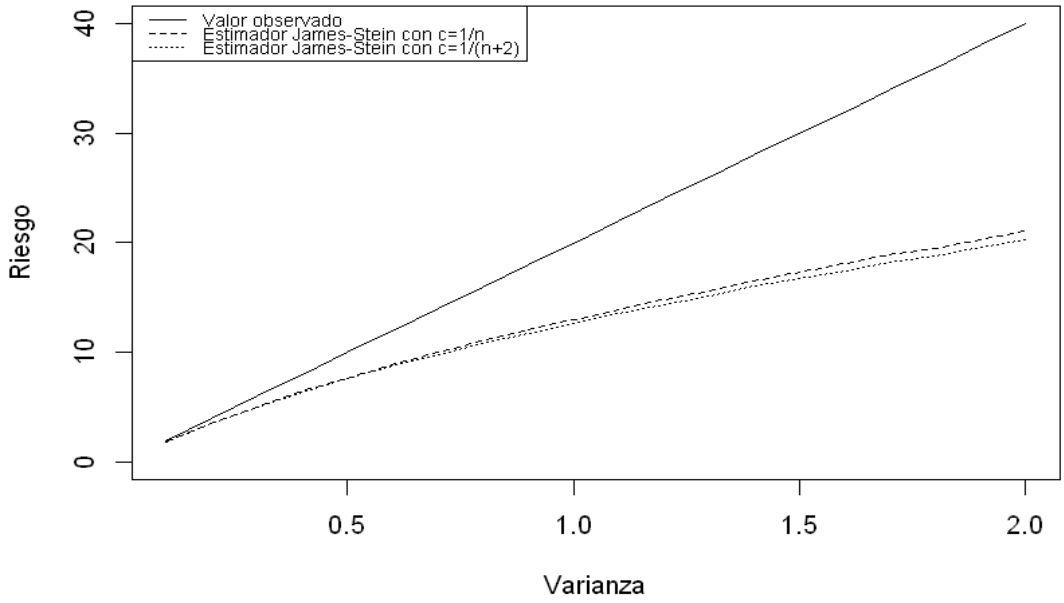


Figura 11: ECM cuando la varianza aumenta

Las siguientes figuras muestran las diferencias entre el caso multivariante de una distribución de Student y el caso normal. La media la hemos puesto a 0 para la Figura 12 y como un múltiplo de $(1, ..., 1)'$ para la Figura 13. La covarianza es la matriz identidad en ambos casos. Vemos que los resultados son muy similares. El ECM de δ^{JS} sigue la misma evolución que el del valor real cuando los grados de libertad aumentan. El grado de mejora es también casi el mismo que en el caso normal.

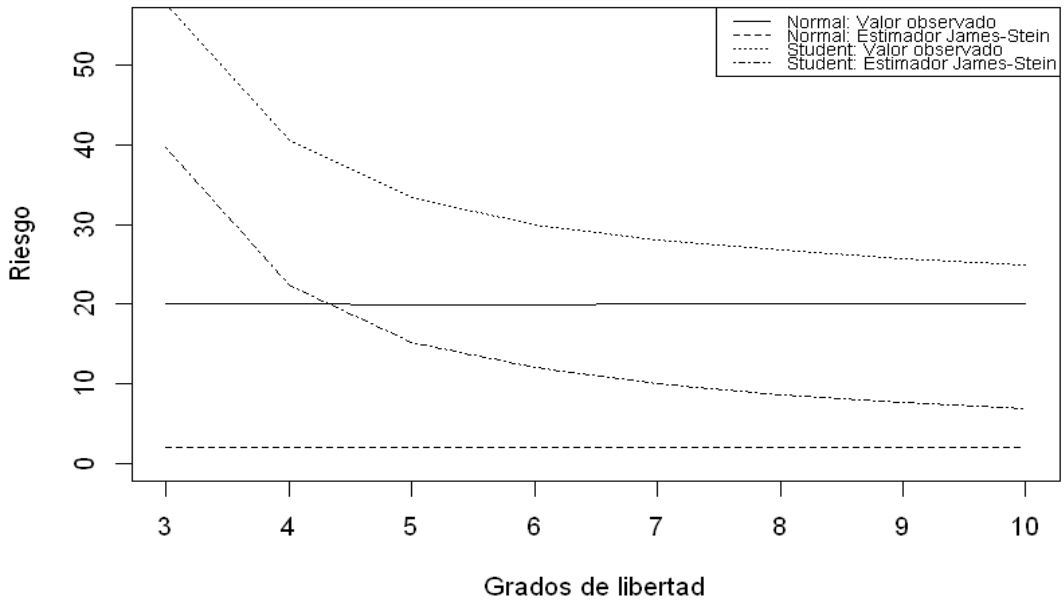


Figura 12: ECM en el caso t-Student y normal conforme los grados de libertad de la t-Student aumentan

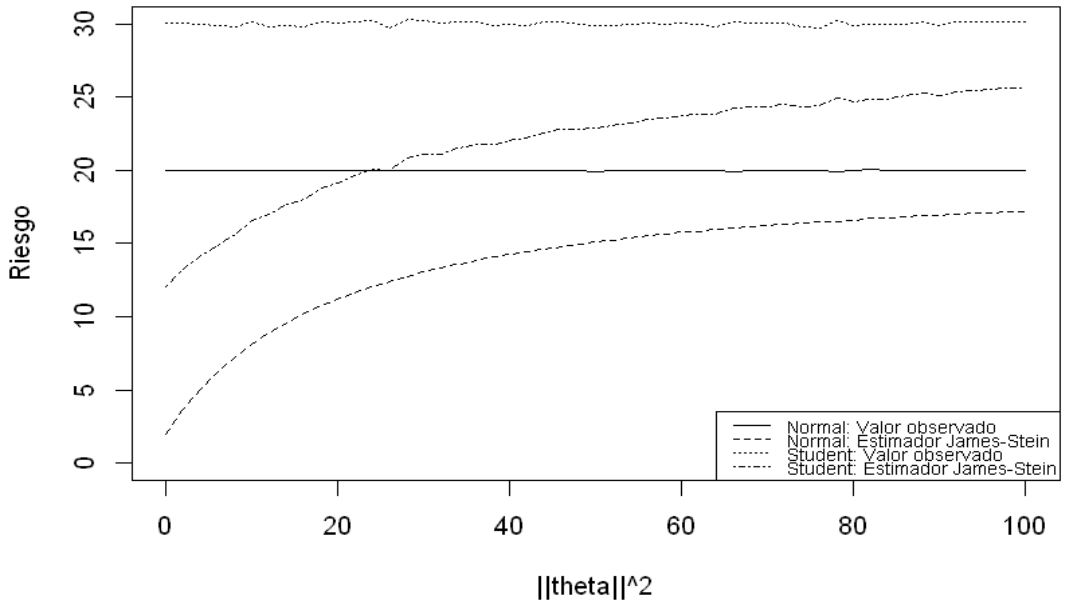
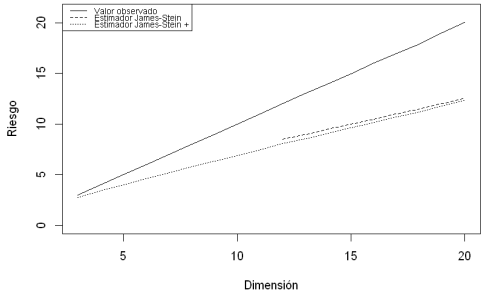
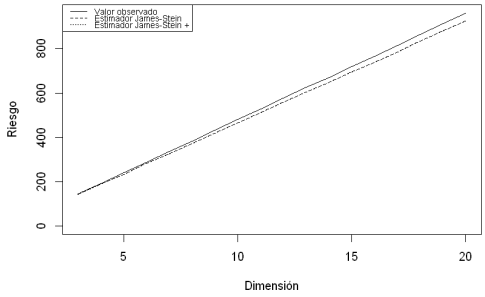


Figura 13: ECM en el caso t-Student y normal conforme la norma de la media aumenta

De igual modo, vemos que para distribuciones de Poisson y ji cuadrado, el estimador de James-Stein encoge.



(a) $X \sim \text{Pois}(1)$



(b) $X \sim \mathcal{X}_6^2$

6. Justificaciones

6.0. Lo que no dice la Paradoja de Stein

El resultado de Stein dice que la terna (media de nacimientos en Cataluña el año 2000, número de espectadores en el mundial de fútbol de 2010, precio del quilo de patatas en China el 2019) es mejorable como estimador del correspondiente vector de parámetros. En la introducción hemos redactado deliberadamente este ejemplo de un modo algo ambiguo para resaltar la sorpresa del resultado.

Es decir, los componentes del vector de medias son admisibles individualmente, pero el vector entero es inadmisible cuando la dimensión es mayor que cierto valor crítico. De manera que no tenemos "mejores" medias sino una "media global" mejor.

Stein empieza describiendo el problema multivariante y luego da un argumento heurístico y geométrico para mostrar que el estimador usual debe ser inadmisible si la dimensión es lo suficientemente grande. Sin embargo, no aclara por qué la dimensión crítica es tres, lo cual ilustramos en la Subsección 6.3.

6.1. Visión Bayesiana

En esta sección, asumimos que la muestra X es de una población con familia paramétrica $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, donde $\Theta \subset \mathbb{R}^p$ para un entero fijo $p \geq 1$.

En la formulación Bayesiana toda cantidad desconocida (como los parámetros del modelo) se trata como una variable aleatoria. Los parámetros tienen una distribución de probabilidad. Antes de observar las variables x, como parte del modelado, se asigna una distribución a priori π a los parámetros θ. Por simplicidad ponemos densidades para todas las distribuciones. La verosimilitud Bayesiana es la densidad de x condicionada a θ. La fórmula de Bayes actualiza la distribución a priori de θ, combinándola con la observación x, dando lugar a la distribución a posteriori.

Teorema 6.1.1. *(Fórmula de Bayes). Dadas $x \sim f(x|\theta)$, $\theta \sim \pi(\theta)$ y la densidad marginal de x. Entonces, la densidad a posteriori de x viene dada por*

$$f(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}, \qquad \text{donde } f(x|\theta)\pi(\theta) = f(x,\theta) = f_\theta(x). \tag{6.1}$$

- *Bayes Empírico.* Los parámetros de la distribución a priori se estiman de los datos. Por ejemplo,

  $$x|\theta \sim \mathcal{N}(\theta, 1),$$
  $$\theta|\tau^2 \sim \mathcal{N}(0, \tau^2).$$

  La distribución marginal de x, $\mathcal{N}(0, \tau^2 + 1)$, se usa para estimar τ.

- *Bayes Jerárquico.* Los parámetros de la distribución a priori son, a su vez, modelados por otra distribución. Por ejemplo,

  $$x|\theta \sim \mathcal{N}(\theta, 1),$$
  $$\theta|\tau^2 \sim \mathcal{N}(0, \tau^2),$$
  $$\tau^2 \sim \mathcal{U}(0, M), \text{ con } M = cte \in (0, \infty).$$

6.1.1. Riesgo de Bayes y estimador de Bayes

Definición 6.1.1. El *riesgo de Bayes* de $\delta : \mathscr{X} \longrightarrow \mathcal{F}$, con $\mathcal{F}$ un conjunto de posible acciones, como

$$\begin{aligned} r(\delta, \pi) &= \int_{\Theta} \mathcal{R}(\delta, \theta)\pi(\theta)d(\theta) \\ &= \int_{\Theta} \left[\int_{\mathscr{X}} L(\delta(x), \theta) f(x|\theta) dx \right] d\pi(\theta) \\ &= \int_{\mathscr{X}} \left[\int_{\Theta} L(\delta(x), \theta) f(\theta|x) d\theta \right] m(x) dx. \end{aligned}$$

Definición 6.1.2. Un estimador es de *Bayes* si minimiza el riesgo de Bayes.

Observación 6.1.1. El estimador que minimiza el valor esperado de la función de pérdida después del evento, $\mathbb{E}(L(\delta(x), \theta)|x)$ también minimiza el riesgo de Bayes y por lo tanto es un estimador de Bayes.

Proposición 6.1.1. *Si x es normal de parámetros $(\theta, \sigma^2 I_p)$ y la distribución a priori de θ es también normal $\sim (\mu, \tau^2 I_p)$, con $\sigma^2, \tau^2 > 0$, y μ conocidos y $m(x) < \infty$. Entonces,*

1. *La marginal $m(x)$ de x es una normal de parámetros $(\mu, (\sigma^2 + \tau^2) I_p)$.*
2. *El estimador de Bayes viene dado por:*

$$\delta_{\pi}(x) = \mu + \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right)(x - \mu). \tag{6.2}$$

Demostración. La primera afirmación se desprende del hecho de que la convolución de $\mathcal{N}_p(0, \sigma^2 I_p)$ y $\mathcal{N}_p(\mu, \tau^2 I_p)$ es $\mathcal{N}_p(\mu, (\sigma^2 + \tau^2) I_p)$, Veamos la segunda afirmación,

$$\begin{aligned} \delta_{\pi}(x) &= \mathbb{E}[\theta|x] \\ &= x + \sigma^2 \frac{\nabla m(x)}{m(x)} \\ &= x + \frac{\sigma^2(-(x - \mu))}{\sigma^2 + \tau^2} \\ &= \frac{\tau^2}{\sigma^2 + \tau^2} x + \frac{\sigma^2}{\sigma^2 + \tau^2} \mu \\ &= \mu + \frac{\tau^2}{\sigma^2 + \tau^2}(x - \mu) \\ &= \mu + \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right)(x - \mu). \end{aligned}$$

□

Muchas veces es conveniente expresar las distribuciones a priori jerárquicamente. Normalmente se toma la primera etapa de la jerarquía como una distribución a priori *conjugada*. En nuestro caso, $\theta \sim \mathcal{N}_p(\mu, \tau^2 I_p)$ es una familia conjugada, ya que su distribución a posteriori también es normal,

$$\theta|x \sim \mathcal{N}_p\left(\frac{\tau^2 x + \sigma^2 \mu}{\sigma^2 + \tau^2}, \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} I_p\right). \tag{6.3}$$

Luego, en la segunda etapa del modelo jerárquico, se pone una distribución a priori sobre τ^2. Suponemos que se puede expresar como $\tau^2 = \sigma^2(1-\lambda)/\lambda$ para $0 < \lambda < 1$. De manera que, $\sigma^2 + \tau^2 = \sigma^2/\lambda$, y

$$\frac{\tau^2 x + \sigma^2 \mu}{\sigma^2 + \tau^2} = (1-\lambda)x + \lambda\mu = \mu + (1-\lambda)(x-\mu).$$

Proposición 6.1.2. *El estimador de Bayes jerárquico es de la forma*

$$\delta_\pi(x) = \mu + \mathbb{E}[(1-\lambda)|x](x-\mu).$$

Demostración. Sea $h(\lambda)$ la densidad de la distribución a priori de la segunda etapa de la jerarquía, obtenemos la densidad marginal

$$m(x) = \int_0^1 \left(\frac{\lambda}{2\pi\sigma^2}\right)^{p/2} exp\left(-\frac{\lambda}{2\sigma^2}\|x-\mu\|^2\right) h(\lambda)d\lambda.$$

Entonces,

$$\begin{aligned}
\delta_\pi(x) &= x + \sigma^2 \frac{\nabla m(x)}{m(x)} \\
&= x - \frac{\displaystyle\int_0^1 \lambda^{p/2+1} exp\left(\frac{\lambda}{2\sigma^2}\|x-\mu\|^2\right) h(\lambda)d\lambda}{\displaystyle\int_0^1 \lambda^{p/2} exp\left(\frac{\lambda}{2\sigma^2}\|x-\mu\|^2\right) h(\lambda)d\lambda}(x-\nu) \\
&= \mu + \mathbb{E}[(1-\lambda)|x](x-\mu).
\end{aligned}$$

□

Por otro lado, si vemos a $\pi(\theta|\tau)$ como un caso concreto de la clase de distribuciones a priori parametrizada por τ, entonces, la marginal de la primera etapa,

$$m(x|\tau) = \int_\Theta f_\theta(x) d\pi(\theta|\tau),$$

la podemos considerar como una función de verosimilitud que depende de las observaciones x y del parámetro τ. De este modo obtenemos los *estimadores de Bayes empíricos.*

Consideramos τ en el sentido frecuentista, y calculamos un estimador de Bayes de la primera etapa sustituyendo el valor estimado λ.

En el caso anterior, la distribución marginal de la primera etapa es

$$x|\tau^2 \sim \mathcal{N}_p(\mu, (\sigma^2 + \tau^2)I_p).$$

Definición 6.1.3. Bajo las suposiciones anteriores, el estimador de Bayes empírico basado en el EMV de τ^2 es

$$\delta^{EB} = \mu + \left(1 - \frac{p\sigma^2}{\|x-\mu\|^2}\right)_+ (x-\mu). \tag{6.4}$$

Demostración. Como μ es constante y conocido, $\|x-\mu\|^2$ es un estadístico canónico. Por lo tanto, el EMV $(\tau^2) = \hat{\tau}^2 = \text{máx}\left(0, \frac{\|x-\mu\|^2}{p} - \sigma^2\right)$. Finalmente,

$$\begin{aligned}\delta^{EB} &= \mu + \frac{\hat{\tau}^2}{\sigma^2+\tau^2}(x-\mu) \\ &= \mu + \left(1 - \frac{p\sigma^2}{\|x-\mu\|^2}\right)_+ (x-\mu). \end{aligned} \tag{6.5}$$

Observamos que es una versión del estimador de la parte positiva del estimador de James-Stein, con $\delta^{EB} \approx \delta_+^{JS}$ cuando $p \longrightarrow \infty$.

Definición 6.1.4. El estimador de Bayes empírico asociado al UMV de $(\sigma^2+\tau^2)^{-1}$, $\frac{p-2}{\|x-\mu\|^2}$, es

$$\delta^{EB} = \mu + \left(1 - \frac{(p-2)\sigma^2}{\|x-\mu\|^2}\right)(x-\mu). \tag{6.6}$$

En este caso, $\delta^{EB} = \delta^{JS}$.

De este modo aclaramos la relación que hay entre los estimadores de James-Stein y la visión empírica Bayesiana.

Observación 6.1.2. Si la varianza a priori, τ^2, tiende a infinito, el estimador de Bayes tenderá a la media muestral. Es decir, cuanto más vaga sea la información a priori, el estimador de Bayes tenderá a dar más peso a la información de la muestra. Por otro lado, si la información a priori es buena, $\sigma^2 > \tau^2$, entonces se pondrá más peso en la media a priori.

6.2. Distribuciones esféricamente simétricas

6.2.1. Condición suficiente

Veremos que podemos limitar la búsqueda de alternativas de $\delta_0 = X$ a la clase de estimadores esféricamente simétricos.

Definición 6.2.1. Un estimador es *esféricamente simétrico* si

$$\delta(X) = \eta(\|X\|)X = [1 - h(\|X\|)]X, \tag{6.7}$$

para ciertas funciones escalares η, h. O equivalente, que para toda transformación ortogonal g, $g \circ \delta \circ g^{-1} = \delta$.

Observación 6.2.1. El estimador usual $\delta_0(X)$ es esféricamente simétrico.

Proposición 6.2.1. *Si estimador esféricamente simétrico $\hat{\delta}$ es admisible en la clase de estimadores esféricamente simétricos, entonces es admisible.*

Demostración. La prueba de esto se basa en la compacidad del grupo ortogonal $\mathcal{O}$, y en la continuidad del problema. Veámoslo por reducción al absurdo.

Suponemos que el estimador δ domina $\hat{\delta}$,

$$\mathcal{R}_\delta(\theta) = \mathbb{E}_\theta[(\delta(X)-\theta)^2] \leq \mathbb{E}_\theta[(\hat{\delta}(X)-\theta)^2]. \tag{6.8}$$

Dada la continuidad de $\mathcal{R}$, obtenemos la desigualdad estricta para al menos un θ de un abierto no vacío U. Ya que $\hat{\delta}$ es esféricamente simétrico, 6.8 se sostiene si reemplazamos δ por $g \circ \delta \circ g^{-1}$, con g ortogonal. De hecho tenemos

$$\mathcal{R}_{g\circ\delta\circ g^{-1}}(\theta) = \mathbb{E}_\theta[\big(g(\delta(g^{-1}X)) - \theta\big)^2] \tag{6.9}$$

$$= \mathbb{E}_\theta[\big(\delta(g^{-1}X) - g^{-1}\theta\big)^2] = \mathcal{R}_\delta(g^{-1}\theta). \tag{6.10}$$

Además, si $\theta \in U$, el conjunto de g para el cual $\mathcal{R}_{g\circ\delta\circ g^{-1}}(\theta) < \mathcal{R}_{\hat{\delta}}(\theta)$ es no vacío. Sea π la medida de probabilidad invariante en $\mathcal{O}$ que asigna las medidas estrictamente positivas a cualquier conjunto abierto no vacío. Entonces,

$$\delta' = \int g \circ \delta \circ g^{-1} d\pi(g) \tag{6.11}$$

es esféricamente simétrico, y por la desigualdad de Jensen tenemos

$$\mathcal{R}_{\delta'}(\theta) \leq \int \mathcal{R}_{g\circ\delta\circ g^{-1}}(\theta) d\mu(g) \leq \mathcal{R}_{\hat{\delta}}(\theta), \tag{6.12}$$

con desigualdad estricta para $\theta \in U$. Por lo tanto $\hat{\delta}$ no es admisible en la clase de estimadores esféricamente simétricos. □

Es decir, δ_0 es inadmisible si y solo si existe un estimador esféricamente simétrico que es mejor.

Brandwein y Strawderman (2012) ampliaron los resultados de Stein en distribuciones esféricamente simétricas con estimadores de la forma $X + ag(X)$, con a constante.

6.2.2. Explicación geométrica

Las distribuciones de los estimadores esféricamente simétricos dependen de la magnitud de θ, no de su dirección, por lo que asumimos $\theta = (\nu, 0, ..., 0)'$ y escribimos $X = (X_1, X'_{(2)})'$, donde $X'_{(2)} \in \mathbb{R}^{p-1}$ es el residuo de X después de proyectar en la dirección de θ. Tomando $R = \|X'_{(2)}\|$, obtenemos

$$Z = (X_1, R) \qquad \text{con } X_1 \sim \mathcal{N}(\nu, 1), \quad R^2 \sim \mathcal{X}^2_{p-1}. \tag{6.13}$$

con X_1 y R independientes. De este modo, expresamos los estimadores esféricamente simétricos en el sistema de coordenadas bidimensional de Z. Además mantiene los riesgos bajo pérdida cuadrática, ya que

$$\|\delta(X) - \theta\| = \|\delta(Z) - (\nu, 0)\|. \tag{6.14}$$

Consideramos $\xi = (\xi_1, \xi_2) = (\nu, \sqrt{p-1})$, el centro "intuitivo" de Z. Como $\mathbb{E}(R^2) = p-1$, tiene sentido tomar $\sqrt{p-1} = \xi_2$.

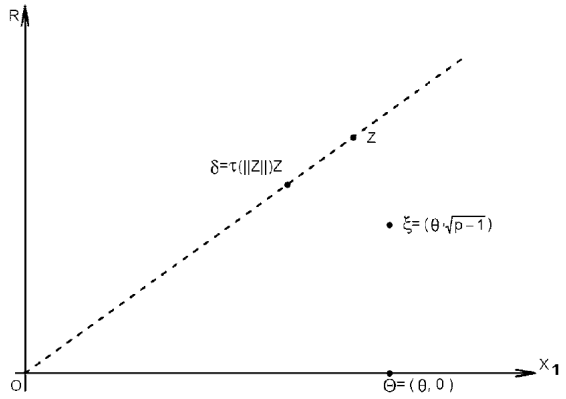


Figura 15: Observación típica en el sistema de coordenadas $Z = (X_1, R)$.

Nos limitamos a estimadores sobre la línea $\overline{OZ}$ y cercanos a θ. Observamos de la imagen que contraer las observaciones hacia el origen a menudo nos será ventajoso.

Observación 6.2.2. Stein (1956) mostró que fijando θ, y a partir de la distribución ji cuadrado no centrada de $\|X\|^2$ o de una aproximación vía series de Taylor, se obtiene

$$\|X\|^2 = \|\theta\|^2 + p + O_p(\sqrt{p}), \qquad \text{si } p \longrightarrow \infty. \tag{6.15}$$

De modo que

$$\|\theta\| = \sqrt{\|X\|^2 - p - O_p(\sqrt{p})} = \|X\| - \frac{p + O_p(\sqrt{p})}{2\|X\|}. \tag{6.16}$$

Entonces, nuestra estimación se encoge tomando como factor de contracción $\dfrac{p + O_p(\sqrt{p})}{\|X\|}$. Incluso con $\dfrac{p + O(1)}{\|X\|}$.

De aquí la motivación a usar estimadores del tipo $\delta_p(X) = \left(1 - \dfrac{p}{\|X\|^2}\right) X$. Y dado que $p \longrightarrow \infty$, la diferencia entre el factor $\dfrac{p}{\|X\|^2}$ y el factor James-Stein es irrelevante.

Teorema 6.2.1. *Dado $Z = (X_1, R)$, los estimadores de la forma*

$$\delta_C(Z) = \left(1 - \frac{C}{\|Z\|^2}\right) Z \tag{6.17}$$

dominan al estimador usual para $0 < C < 2(p-2)$, en particular también para $p \geq 3$ y $C = p - 1$. Y son óptimos para $C = p - 2$, obteniendo así el estimador de James-Stein.

Demostración. Partiendo de la Figura 16, aproximamos los cálculos de los riesgos a partir de los valores ξ_+, ξ_- definidos como

$$\xi_\pm = \left(\nu \pm 1, \sqrt{p-1}\right).$$

Condicionaremos a $Z = \xi_\pm$, ya que estos puntos son cercanos a ξ, y además exhiben la variación estocástica típica en la dirección de $\theta = (\nu, 0)$, dado que su media y ECM en esa dirección coinciden con los de la distribución completa.

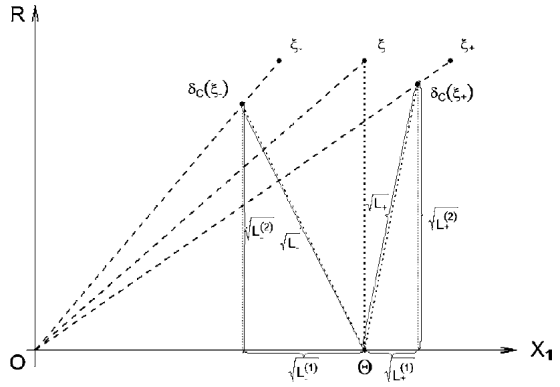


Figura 16: Los valores de $\xi_\pm$ y sus estimaciones respectivas.

Desglosamos los cálculos en dos componentes, las correspondientes a las direcciones determinadas por las coordenadas de Z. Sea $L_\pm$ el error cuadrático de una observación respecto a $\xi_\pm$,

$$\begin{aligned} L_\pm &= \left[\left(1 - \frac{C}{\|\xi_\pm\|^2}\right)(\nu \pm 1) - \nu\right]^2 + \left[\left(1 - \frac{C}{\|\xi_\pm\|^2}\right)\sqrt{p-1} - 0\right]^2 \\ &= \left[\pm 1 - \frac{C}{\|\xi_\pm\|^2}(\nu \pm 1)\right]^2 + \left(1 - \frac{C}{\|\xi_\pm\|^2}\right)^2 (p-1) \\ &= L_\pm^{(1)} + L_\pm^{(2)}. \end{aligned}$$

Entonces,

$$\mathcal{R}_{|\xi_\pm} \equiv \mathcal{R}_{|\xi_\pm}(\delta_C, \theta) = \frac{1}{2}(L_+^{(1)} + L_-^{(1)}) + \frac{1}{2}(L_+^{(2)} + L_-^{(2)}) = \mathcal{R}_{|\xi_\pm}^{(1)} + \mathcal{R}_{|\xi_\pm}^{(2)}. \tag{6.18}$$

Para δ_0 la diferencia de coordenadas en los riesgos condicionales, es 1 y $p-1$, respectivamente. Por lo tanto, las diferencias de riesgos respecto a δ_C son

$$1 - \mathcal{R}_{|\xi_\pm}^{(1)} = \frac{1}{2}\left(\frac{2C(\nu+1)}{\|\xi_+\|^2} - \frac{C^2(\nu+1)^2}{\|\xi_+\|^4} - \frac{2C(\nu-1)}{\|\xi_-\|^2} - \frac{C^2(\nu-1)^2}{\|\xi_-\|^4}\right)$$

y

$$(p-1) - \mathcal{R}_{|\xi_\pm}^{(2)} = \frac{1}{2}(p-1)\left(\left(\frac{2C}{\|\xi_+\|^2} + \frac{2C}{\|\xi_-\|^2}\right) - \left(\frac{C^2}{\|\xi_+\|^4} + \frac{C^2}{\|\xi_-\|^4}\right)\right).$$

Reorganizando,

$$\begin{aligned} \Delta_{|\xi_\pm} &\coloneqq p - \mathcal{R}_{|\xi_\pm} \\ &= C\nu\left(\frac{1}{\|\xi_+\|^2} - \frac{1}{\|\xi_-\|^2}\right) + Cp\left(\frac{1}{\|\xi_+\|^2} - \frac{1}{\|\xi_-\|^2}\right) \\ &- \frac{1}{2}C^2\left(\frac{(\nu+1)^2 + p - 1}{\|\xi_+\|^4} + \frac{(\nu-1)^2 + p - 1}{\|\xi_-\|^4}\right) \\ &= C\nu\left(\frac{1}{\|\xi_+\|^2} - \frac{1}{\|\xi_-\|^2}\right) + \left(Cp - \frac{1}{2}C^2\right)\left(\frac{1}{\|\xi_+\|^2} + \frac{1}{\|\xi_-\|^2}\right) \end{aligned} \tag{6.19}$$

ya que $\|\xi_\pm\|^2 = (\nu \pm 1)^2 + p - 1$. Queremos $\Delta > 0$, para que δ_C domine a δ_0. Consideramos dos opciones.

Si $\|\xi_+\|^2 = \|\xi_-\|^2$, entonces 6.19 sería positivo para cualquier $0 < C < 2p$. En particular, tomando $p \geq 2$ sería positivo para $C = p - 1$ (incluso lo sería con $p = 1$). No tiene sentido como solución.

Suponemos pues $\|\xi_+\|^2 > \|\xi_-\|^2$ de modo que $\dfrac{1}{\|\xi_+\|^2} - \dfrac{1}{\|\xi_-\|^2} < 0$. En más detalle,

$$\frac{1}{\|\xi_+\|^2} - \frac{1}{\|\xi_-\|^2} = \frac{\|\xi_-\|^2 - \|\xi_+\|^2}{\|\xi_+\|^2\|\xi_-\|^2} = -4\frac{\nu}{\|\xi_+\|^2\|\xi_-\|^2}$$
$$\frac{1}{\|\xi_+\|^2} + \frac{1}{\|\xi_-\|^2} = \frac{\|\xi_-\|^2 + \|\xi_+\|^2}{\|\xi_+\|^2\|\xi_-\|^2} = 2\frac{\nu^2 + p}{\|\xi_+\|^2\|\xi_-\|^2}.$$

Por lo que la diferencia con riesgos condicionales para $p \geq 2$ es

$$\begin{aligned}\Delta_{|\xi_\pm} &= \frac{2}{\|\xi_+\|^2\|\xi_-\|^2}\left(\left(C(p-2) - \frac{C^2}{2}\right)\nu^2 + \left(Cp - \frac{C^2}{2}\right)p\right) \\ &> \frac{2(\nu^2+p)}{\|\xi_+\|^2\|\xi_-\|^2}\left(C(p-2) - \frac{C^2}{2}\right). \end{aligned} \tag{6.20}$$

□

Observación 6.2.3. En 6.20 los tres términos de la fracción principal son aproximadamente iguales; esto es, $\nu^2 + p \approx \|\xi_+\|^2 \approx \|\xi_-\|^2$.

Por lo tanto, la diferencia en los riesgos, $\Delta = \mathcal{R}(\delta_0, \theta) - \mathcal{R}(\delta_C, \theta)$, se aproxima bien como

$$\Delta \approx \frac{2}{\|\theta\|^2 + p}\left(C(p-2) - \frac{C^2}{2}\right). \tag{6.21}$$

□

Observación 6.2.4. La calidad de esta aproximación mejora conforme $\|\theta\| \longrightarrow \infty$.

Para permitir cálculos más precisos para $\|\theta\|$ grandes, reemplazamos δ_C con el estimador

$$\delta_{C;a} = \left(1 - \frac{C}{a + \|X\|^2}\right)X.$$

Y luego desarrollando series de Taylor (ver Stein, 1956) obtenemos

$$\mathcal{R}(\delta_0, \theta) - \mathcal{R}(\delta_{C;a}, \theta) = \frac{2}{a + \|\theta\|^2}\left(C(p-2) - \frac{C^2}{2}\right) + o\left(\frac{1}{a + \|\theta\|^2}\right). \tag{6.22}$$

De ello deducimos que en este caso δ_0 es inadmisible.

Observación 6.2.5. Si consideramos solo la configuración de distribución normal, entonces

$$\Delta = \mathcal{R}(\delta_0, \theta) - \mathcal{R}(\delta_C, \theta) = \mathbb{E}_\theta\left(\frac{1}{\|X\|^2}\right)\left(C(p-2)\frac{C^2}{2}\right). \tag{6.23}$$

De este modo,

$$\mathbb{E}_\theta\left(\frac{1}{\|X\|^2}\right) \approx \frac{1}{\mathbb{E}_\theta(\|X\|^2)} = \frac{1}{\|\theta\|^2 + p}, \tag{6.24}$$

siendo la aproximación bastante cercana excepto cuando $\|\theta\|$ es pequeña. Por lo tanto, la aproximación heurística en 6.20 y 6.21 está bastante cerca de la verdad. Esto valida la idea de aproximar la diferencia en riesgos por la diferencia condicional dada $\theta = \xi_+, \xi_-$.

6.3. Relación de estimación de Stein con teoría de probabilidad

Después de ver que la media empírica es admisible en dimensiones 1 y 2, pero no lo es partir de dimensión 3, la pregunta obvia es: ¿Qué pasa con el 3?

Existe otro resultado en teoría de Probabilidad sobre una propiedad que es cierta en dimensiones 1 y 2, pero no lo es a partir de dimensión 3. En el caso discreto, es el teorema de George Polya de recurrencia del paseo aleatorio en el retículo entero de dimensión p, $\mathbb{Z}^p$. Este teorema afirma que este paseo aleatorio es *recurrente* para $p = 1, 2$ y *transitorio* para $p \geq 3$ (ver Novak, 2014).

No es mera coincidencia: ambas condiciones, la de admisibilidad de Stein y la de recurrencia están relacionadas. Brown (1971) establece que: Dado X vector aleatorio normal p-variante $\sim \mathcal{N}_p(\theta, I_p)$, para cada estimador invariante $\hat{\delta} \equiv \hat{\delta}(X)$ del vector θ con riesgo cuadrático finito, existe un proceso de difusión en el espacio euclídeo p-dimensional (que construye explícitamente) que es recurrente si, y solo si, el estimador es admisible.

La media local del proceso de difusión es la diferencia $\hat{\delta} - X$ y su matriz de covarianzas es $2\, I_p$, así al EMV le corresponde el movimiento Browniano (escalado).

(Ver también Bhattacharya, 1978)

7. El efecto contrario

Suponemos ahora que $X = Y + \delta \in \mathbb{R}^p$ es el vector aleatorio observado. Siendo δ un parámetro desconocido de posición e Y el vector aleatorio no observado absolutamente continuo. Además, asumimos que $Y \equiv X - \delta$ es direccionalmente simétrico, es decir que $\overrightarrow{Y} \overset{d}{=} -\overrightarrow{Y}$, donde $\overrightarrow{Y} = \dfrac{Y}{\|Y\|}$ es el vector unitario en la dirección de Y.

Consideramos los estimadores de contracción δ de la siguiente forma

$$\hat{\delta}_\gamma \equiv \hat{\delta}_\gamma(X; \delta_0) = \gamma(X - \delta_0) \cdot (X - \delta_0) + \delta_0, \tag{7.1}$$

donde $\gamma \equiv \gamma(X - \delta_0) \in [0, 1)$ y δ_0 es cualquier punto fijo de $\mathbb{R}^p$ hacia el cual queremos contraer.

Fijamos δ y δ_0, y consideramos $B_1 \equiv B_1(\|\delta - \delta_0\|; \delta_0) \subset \mathbb{R}^p$ la bola de radio $\|\delta - \delta_0\|$ centrada en δ_0.
Tomamos H como la mitad del espacio delimitado por el hiperplano ∂H tangente a B_1 en δ. De este modo,

$$Pr_\delta\left[\|X - \delta_0\| > \|\delta - \delta_0\| | \delta_0\right] = Pr_\delta\left[X \in B_1^c | \delta_0\right] > Pr_\delta\left[X \in H | \delta_0\right] = \frac{1}{2}. \tag{7.2}$$

Además, suponiendo:

- δ_0 fijo, es independiente de X.
- $Y \equiv X - \delta$ es esféricamente simétrico.
- $\dfrac{\|\delta_0 - \delta\|}{\|Y\|} = \dfrac{\|\delta_0 - \delta\|}{\|X - \delta\|} = o(\sqrt{p})$ en probabilidad.

Obtenemos,

$$\lim_{p \longrightarrow \infty} Pr_\delta\left[\|X - \delta_0\| > \|\delta - \delta_0\| | \delta_0\right] \equiv \lim_{p \longrightarrow \infty} Pr_\delta\left[X \in B_1^c\right] = 1. \tag{7.3}$$

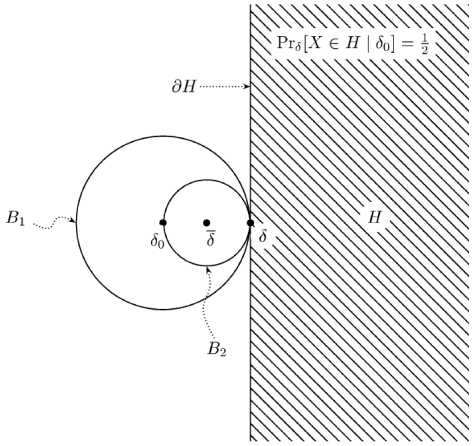


Figura 17: Bolas B_1 y B_2.

Por lo tanto, $\|X - \delta_0\|$ es una sobreestimación de $\|\delta - \delta_0\|$, así que un estimador de la forma $\hat{\delta}_\gamma$ debería ser preferible a $X - \delta_0$.

Supongamos que $\tilde{\gamma} \equiv \tilde{\gamma}(X - \delta_0, \delta - \delta_0)$ depende de δ y tomamos $B_2 \equiv B_2(\|\delta - \delta_0\|; \bar{\delta})$, la bola de radio $\frac{1}{2}\|\delta - \delta_0\|$ y centro $\bar{\delta} \equiv \frac{1}{2}(\delta_0 + \delta)$. Dado que $B_2^c \supset B_1^c$,

$$Pr_\delta \left[X \in B_2^c | \delta_0 \right] > \frac{1}{2} \tag{7.4}$$

y, bajo las suposiciones anteriores,

$$\lim_{p \longrightarrow \infty} Pr_\delta \left[X \in B_2^c \right] = 1. \tag{7.5}$$

Si δ fuera conocido, entonces cierto factor de contracción γ aplicado a $X - \delta_0$ acercaría X a δ.

Asumimos $Y \sim \mathcal{N}_p(0, \sigma^2 I_p)$, con σ^2 conocida, de manera que $X \sim \mathcal{N}_p(\delta, \sigma^2 I_p)$.

En este caso, el estimador de James-Stein para δ viene dado por

$$\hat{\delta}^{JS} \equiv \hat{\delta}^{JS}(X; \delta_0) = \left(1 - \frac{\sigma^2 (p-2)}{\|X - \delta_0\|^2} \right) (X - \delta_0) + \delta_0, \tag{7.6}$$

Junto con la versión positiva de éste, tienen la propiedad de que para $p \geq 3$ dominan a X bajo el error cuadrático medio y el criterio de cercanía de Pitman (PC). Es decir,

$$\mathbb{E}_\delta \left[\|\hat{\delta}_+^{JS}(X; \delta_0)\|^2 \middle| \delta_0 \right] < \mathbb{E}_\delta \left[\|\hat{\delta}^{JS}(X; \delta_0)\|^2 \middle| \delta_0 \right] < \mathbb{E}_\delta \left[\|X - \delta\|^2 \right] \equiv p\sigma^2. \tag{7.7}$$

Y

$$\begin{aligned} Pr_\delta \left[\|\hat{\delta}_{JS}^+(X; \delta_0) - \delta\| < \|X - \delta\| \middle| \delta_0 \right] &> Pr_\delta \left[\|\hat{\delta}_{JS}(X; \delta_0) - \delta\| < \|X - \delta\| \middle| \delta_0 \right] \\ &= Pr \left[\mathcal{X}_p^2 \left(\frac{\|\delta - \delta_0\|^2}{4\sigma^2} \right) \geq \frac{\|\delta - \delta_0\|^2}{4\sigma^2} + \frac{p-2}{2} \right] \\ &> \frac{1}{2}, \end{aligned} \tag{7.8}$$

donde $\mathcal{X}_p^2(\eta)$ es una variable aleatoria con ley una distribución ji cuadrado no centrada de p grados de libertad y parámetro de no centralidad η.

Notamos especialmente que:

1. Las mejoras que nos ofrecen los estimadores James-Stein son grandes, sobre todo cuando $p \longrightarrow \infty$. Si $\|\delta - \delta_0\| = o(p)$ con σ^2 fijada, entonces tanto para $\hat{\delta} = \hat{\delta}^{JS}$ como para $\hat{\delta} = \hat{\delta}_+^{JS}$ tenemos,

$$Pr_\delta \left[\|\hat{\delta} - \delta\| < \|X - \delta\| \right] \overset{p \to \infty}{\longrightarrow} 1.$$

2. La dominancia de $\hat{\delta}^{JS}$ y $\hat{\delta}_+^{JS}$ respecto a X bajo ECM y PC se sostiene incluso cuando la media, δ, está lejos del objetivo de contracción δ_0.

Observación 7.0.1. De las dos propiedades (1.) y (2.), la segunda es la más sorprendente. No es difícil construir estimadores que satisfagan (1.). Por ejemplo, el estimador de Bayes respecto a una distribución a priori normal centrada en δ_0. De todos modos, este estimador de Bayes no satisface (2.), la diferencia se deriva de que el estimador de Bayes tendrá un factor de contracción fijo mientras que el de James-Stein y el positivo son adaptativos.

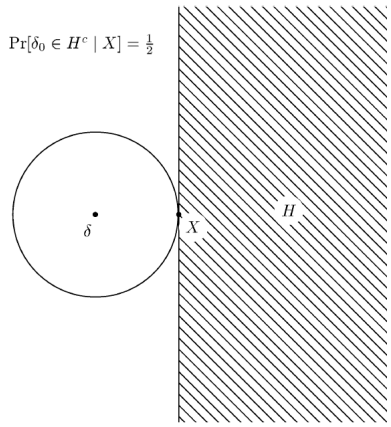


Figura 18: El complementario de H es el conjunto (7.9).

Así pues, suponemos que la posición de $X \sim \mathcal{N}_p(\delta, \sigma^2 I_p)$ es conocida y $p \geq 3$.

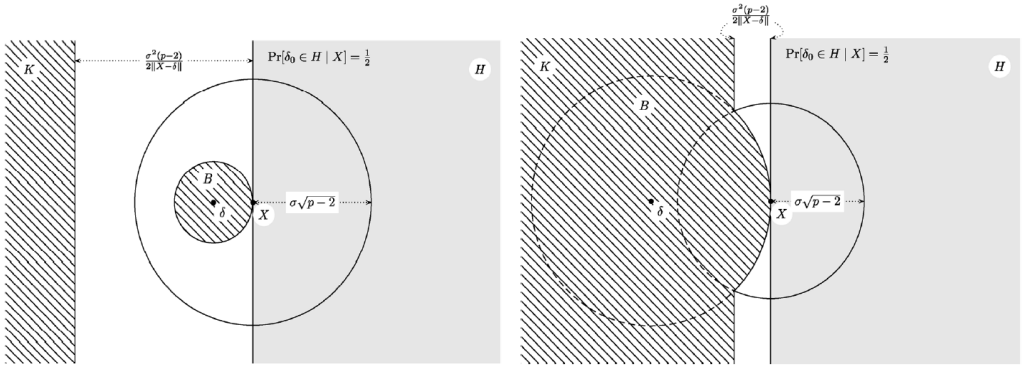
Consideramos el conjunto

$$H^c = \{\delta_0 \mid \exists \tilde{\gamma} \in [0,1) \ni \|\hat{\delta}_{\tilde{\gamma}} - \delta\| < \|X - \delta\|\} \tag{7.9}$$

donde $\tilde{\gamma} \equiv \tilde{\gamma}(X - \delta_0, X - \delta)$ depende de δ.

Dada que $Pr[\delta_0 \in H^c \mid X] = \dfrac{1}{2}$, contraer hacia un δ_0 elegido aleatoriamente tiene como máximo un 50 % de probabilidades de mover X más cerca de δ_0.

Para valores representativos de δ, el conjunto de todos los δ_0 tales que $\hat{\delta}_+^{JS}(X; \delta_0)$ se encuentra más cerca de δ que la observación X tenemos los siguientes dos casos.



(a) $\|X - \delta\| < \dfrac{\sigma}{2}\sqrt{p-2}$ (b) $\|X - \delta\| > \dfrac{\sigma}{2}\sqrt{p-2}$

Figura 19: La región rayada es el conjunto $\{\delta_0 | \|\hat{\delta}_{JS}^{+}(X; \delta_0) - \delta\| < \|X - \delta\|\}$.

Por la simetría direccional de δ_0 sobre X, obtenemos de las figuras

$$Pr\left[\|\hat{\delta}_+^{JS}(X; \delta_0) - \delta\| > \|X - \delta\| \middle| X\right] > Pr\left[\delta_0 \in H | X\right] = \frac{1}{2}. \tag{7.10}$$

Si δ_0 es distribuido simétricamente sobre X entonces tenemos que

$$\mathbb{E}\left[\hat{\delta}_+^{JS}(X;\delta_0)\Big| X\right] = X, \tag{7.11}$$

y por la desigualdad de Jensen,

$$\mathbb{E}\left[\|\hat{\delta}_+^{JS}(X;\delta_0)-\delta\|^2\Big| X\right] > \mathbb{E}\left[\|X-\delta\|^2\big| X\right] \equiv p\sigma^2, \qquad \forall\delta\in\mathbb{R}^p. \tag{7.12}$$

Además, bajo las suposiciones adicionales anteriores,

$$\lim_{p\longrightarrow\infty} Pr_\delta\left[\|\hat{\delta}_+^{JS}(X;\delta_0)-\delta\| > \|X-\delta\|\right] = 1. \tag{7.13}$$

Bajo cualquier distribución de probabilidad de δ_0, 7.7 sostiene que

$$\mathbb{E}_\delta\left[\|\hat{\delta}_+^{JS}(X;\delta_0)-\delta\|^2\right] < \mathbb{E}_\delta\left[\|X-\delta\|^2\right] \equiv p\sigma^2, \qquad \forall\delta\in\mathbb{R}^p, \tag{7.14}$$

mientras que 7.12 sostiene que

$$\mathbb{E}_\delta\left[\|\hat{\delta}_+^{JS}(X;\delta_0)-\delta\|^2\right] > \mathbb{E}_\delta\left[\|X-\delta\|^2\right] \equiv p\sigma^2, \qquad \forall\delta\in\mathbb{R}^p. \tag{7.15}$$

Así pues, ¿estamos contrayendo o no? En realidad, no se ha producido ninguna contradicción formal.

Las probabilidades y esperanzas que aparecen en 7.7, 7.8, 7.10 y 7.12 son condicionadas con diferentes variables condicionantes.

Además, las distribuciones conjuntas de (X,δ_0) en 7.14 y 7.15 son diferentes, sus densidades son $f_\delta(X)f(\delta_0)$ y $f_\delta(X)f(\delta_0|X)$, respectivamente. En el primero, X y δ_0 son independientes mientras que en el segundo, δ_0 es dependiente.

El efecto inverso de Stein es pues tan real como el propio efecto de Stein original, ambos son manifestaciones de la fuerte curvatura de las esferas en el espacio euclidiano multidimensional.

Las figuras y los resultados en 7.10, 7.12 y 7.13 muestran que sin algunos conocimientos previos de la ubicación de δ, no debemos encoger X.

Si en vez de elegir el objetivo de contracción δ_0 con información previa confiable lo tomamos basándonos en los datos X, la robustez minimax/Bayesiana de la propiedad (2.) del estimador JS se pierde y ya no tenemos garantizado que encoger no sea dañino en promedio.

8. Conclusiones

En este trabajo hemos hecho una revisión histórica de la paradoja de Stein así como de los estimadores que se propusieron para mostrar la inadmisibilidad en dimensiones mayores que 2. Además, hemos aportado varias justificaciones a este fenómeno así como gráficos para ilustrarlo.

Mostramos también que un estimador de contracción es tan bueno, pero no mejor, que la información previa en la que se basa. Sin información a priori confiable es probable que la contracción disminuya la precisión de la estimación. Es por ello que este fenómeno es susceptible a malas interpretaciones.

Asimismo, si el problema de estimación estadística es verdaderamente invariante bajo traslaciones, entonces deberíamos usar el mejor estimador invariante, normalmente X.

Más allá de lo especificado en este trabajo, podríamos ampliar los resultados a otras distribuciones. Y también considerar otras funciones de pérdida que no sean el error cuadrático medio.

Los estimadores James-Stein y derivados suelen ser presentados como una idea sorprendente bastante difícil de aceptar y contrarios al pensamiento intuitivo. Es por este motivo que los argumentos presentados en el proyecto aportan luz al tema y permiten seguir de una manera estructurada y ordenada la construcción de estos estimadores y contrastarlos al estimador usual.

Referencias

[1] Bhattacharya, R.N. (1978). Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *The Annals of Probability, 6*(4), 541-553.

[2] Brandwein, A. & Strawderman, W. (2012). Stein Estimation for Spherically Symmetric Distributions: Recent Developments. *Statistical Science, 27*(1), 11-23.

[3] Brown L.D. (1971). Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems. *The Annals of Mathematical Statistics, 42*(3), 855-903.

[4] Brown L.D. & Zhao L. (2012). A Geometrical Explanation to Stein Shrinkage. *Statistical Science, 27*(1), 24-30.

[5] Casella, G. & Berger, R.L. (2002). *Statistical Inference* (2nd ed.). Thomson Learning.

[6] Corcuera, J.M. (2019). Estadística (Notas de clase).

[7] DeGroot, M.H. (1986). A Conversation with Charles Stein. *Statistical Science, 1*(4), 454-462.

[8] Demaret, T. (2019). *About Stein's estimators: the original result and extensions.* (Trabajo de final de máster, University of Liège). http://hdl.handle.net/2268.2/6981

[9] De Wet, T. (2003). Statistics in the Fifties. *Presidential Address on the 50th Anniversary of the South African Statistical Association.*

[10] Efron B. (1992). Introduction to James and Stein (1961): Estimation with Quadratic Loss. En: S. Kotz S. y N.L. Johnson (ed.), *Breakthroughs in Statistics.* Springer Series in Statistics (Perspectives in Statistics). Springer.

[11] Efron, B. (2010). Empirical Bayes and the James-Stein Estimator. En: *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Institute of Mathematical Statistics Monographs, p. 1-14). Cambridge: Cambridge University Press.

[12] Efron, B. & Morris, C. (1977). Stein's Paradox in Statistics. *Scientific American, 236*(5), 119-127.

[13] Fourdrinier, D., Strawderman, W.E. & Wells, M.T. (2018) Decision Theory Preliminaries. En: *Shrinkage Estimation.* Springer Series in Statistics. Springer, Cham.

[14] Hodges J.L y Lehmann E.L. (1951). Some Applications of the Cramér-Rao Inequality. En J. Neyman, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (p. 13-22). University of California Press.

[15] Kubota, T. (2016, diciembre 1). Charles M. Stein, extraordinary statistician and anti-war activist, dies at 96. *Stanford News.* 08 de junio de 2021. https://news.stanford.edu/2016/12/01/charles-m-stein-extraordinary-statistician-anti-war-activist-dies-96/

[16] Leong, Y.K. (2003). Charles Stein: The Invariant, the Direct and the "Pretentious": An interview of Charles Stein. *Newsletter of the American Mathematical Institute, NUS.* 16-19. https://ims.nus.edu.sg/wp-content/uploads/2020/07/imprints-2-2003.pdf

[17] Novak, J.(2014). Pólya's Random Walk Theorem. *The American Mathematical Monthly, 121*(8), 711-716.

[18] Shao J. (2003). Estimation in Parametric Models. En: *Mathematical Statistics*. Springer Texts in Statistics. Springer.

[19] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceeding of the fourth Berkeley symposium on Mathematical statistics and Probability* (Vol. 1, p.197-206). University of California Press.

[20] Stein, C. (1962). Discussion on Professor Stein's Paper. *Royal Statistical Society, 24* 285-296.

[21] Stein, C. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics, 9*(6), 1135-1151.

[22] Stein C. & James W. (1961). Estimation with Quadratic Loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol.1, p. 361-379). University of California Press.

[23] Perlman, M. & Chaudhuri S. (2012). Reversing the Stein Effect. *Statistical Science, 27*(1), 135-143.