Abundancia vs Esparsidad

Liliana Forzani
FIQ (UNL-CONICET)

Regresión

Estudio de la distribución condicional de Y (dependiente o respuesta) dado $\mathbf{X}$ (predictores), es decir $(Y|\mathbf{X})$.

Ejemplos:

- Dadas las alturas de la madre y el padre ($\mathbf{X}$) queremos predecir la altura del hijo (Y).
- Dado un fragmento de sonido ($\mathbf{X}$) queremos identificar (automáticamente): ¿es un ave, un auto o un avión (Y)?

Regresión lineal

Regresión lineal

- Modelo teórico: $Y|\mathbf{X} = \beta^T \mathbf{X} + \epsilon$, $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^p$, p fijo, $\epsilon \sim N(0, \sigma^2)$.

Regresión lineal

- Modelo teórico: $Y|\mathbf{X} = \beta^T \mathbf{X} + \epsilon$, $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^p$, p fijo, $\epsilon \sim N(0, \sigma^2)$.
- ¿β? En población $\beta = \Sigma^{-1}\Sigma_{XY}$ con $\Sigma = \mathrm{var}(\mathbf{X})$ y $\Sigma_{XY} = \mathrm{cov}(\mathbf{X}, Y)$.

Regresión lineal. $n > p$ y p fijo

- Modelo teórico: $Y|\mathbf{X} = \beta^T \mathbf{X} + \epsilon$, $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^p$, p fijo, $\epsilon \sim N(0, \sigma^2)$.
- ¿β? En población $\beta = \Sigma^{-1}\Sigma_{XY}$ con $\Sigma = \mathrm{var}(\mathbf{X})$ y $\Sigma_{XY} = \mathrm{cov}(\mathbf{X}, Y)$.
- Dado un conjunto de datos $(\mathbf{X}_i, Y_i) \in \mathbb{R}^{p+1}$, $i = 1, \ldots, n$, $\mathbb{X} \in \mathbb{R}^{n \times p}$, $\mathbb{Y} \in \mathbb{R}^{n \times 1}$ siguiendo el modelo:

Regresión lineal. $n > p$ y p fijo

- Modelo teórico: $Y|\mathbf{X} = \beta^T\mathbf{X} + \epsilon$, $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^p$, p fijo, $\epsilon \sim N(0, \sigma^2)$.
- ¿β? En población $\beta = \Sigma^{-1}\Sigma_{XY}$ con $\Sigma = \mathrm{var}(\mathbf{X})$ y $\Sigma_{XY} = \mathrm{cov}(\mathbf{X}, Y)$.
- Dado un conjunto de datos $(\mathbf{X}_i, Y_i) \in \mathbb{R}^{p+1}$, $i = 1, \ldots, n$, $\mathbb{X} \in \mathbb{R}^{n\times p}$, $\mathbb{Y} \in \mathbb{R}^{n\times 1}$ siguiendo el modelo:
  - Estimar β: $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$, con $\widehat{\Sigma} = \frac{1}{n}\mathbb{X}^T\mathbb{X}$ y $\widehat{\Sigma}_{XY} = \frac{1}{n}\mathbb{X}^T\mathbb{Y}$.

Regresión lineal. $n > p$ y p fijo

- Modelo teórico: $Y|\mathbf{X} = \beta^T\mathbf{X} + \epsilon$, $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^p$, p fijo, $\epsilon \sim N(0, \sigma^2)$.
- ¿β? En población $\beta = \Sigma^{-1}\Sigma_{XY}$ con $\Sigma = \mathrm{var}(\mathbf{X})$ y $\Sigma_{XY} = \mathrm{cov}(\mathbf{X}, Y)$.
- Dado un conjunto de datos $(\mathbf{X}_i, Y_i) \in \mathbb{R}^{p+1}$, $i = 1, \ldots, n$, $\mathbb{X} \in \mathbb{R}^{n\times p}$, $\mathbb{Y} \in \mathbb{R}^{n\times 1}$ siguiendo el modelo:
  - Estimar β: $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$, con $\widehat{\Sigma} = \frac{1}{n}\mathbb{X}^T\mathbb{X}$ y $\widehat{\Sigma}_{XY} = \frac{1}{n}\mathbb{X}^T\mathbb{Y}$.
  - Analizar residuales para validar el modelo.

Regresión lineal. $n > p$ y p fijo

- Modelo teórico: $Y|\mathbf{X} = \beta^T\mathbf{X} + \epsilon$, $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^p$, p fijo, $\epsilon \sim N(0, \sigma^2)$.
- ¿β? En población $\beta = \Sigma^{-1}\Sigma_{XY}$ con $\Sigma = \mathrm{var}(\mathbf{X})$ y $\Sigma_{XY} = \mathrm{cov}(\mathbf{X}, Y)$.
- Dado un conjunto de datos $(\mathbf{X}_i, Y_i) \in \mathbb{R}^{p+1}$, $i = 1, \ldots, n$, $\mathbb{X} \in \mathbb{R}^{n\times p}$, $\mathbb{Y} \in \mathbb{R}^{n\times 1}$ siguiendo el modelo:
  - Estimar β: $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$, con $\widehat{\Sigma} = \frac{1}{n}\mathbb{X}^T\mathbb{X}$ y $\widehat{\Sigma}_{XY} = \frac{1}{n}\mathbb{X}^T\mathbb{Y}$.
  - Analizar residuales para validar el modelo.
  - Acompañar la estimación de β con una región de confianza que involucra $(\mathbb{X}^T\mathbb{X})^{-1}$.

Regresión lineal. $n > p$ y p fijo

- Modelo teórico: $Y|\mathbf{X} = \beta^T\mathbf{X} + \epsilon$, $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^p$, p fijo, $\epsilon \sim N(0, \sigma^2)$.
- ¿β? En población $\beta = \Sigma^{-1}\Sigma_{XY}$ con $\Sigma = \mathrm{var}(\mathbf{X})$ y $\Sigma_{XY} = \mathrm{cov}(\mathbf{X}, Y)$.
- Dado un conjunto de datos $(\mathbf{X}_i, Y_i) \in \mathbb{R}^{p+1}$, $i = 1, \ldots, n$, $\mathbb{X} \in \mathbb{R}^{n\times p}$, $\mathbb{Y} \in \mathbb{R}^{n\times 1}$ siguiendo el modelo:
  - Estimar β: $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$, con $\widehat{\Sigma} = \frac{1}{n}\mathbb{X}^T\mathbb{X}$ y $\widehat{\Sigma}_{XY} = \frac{1}{n}\mathbb{X}^T\mathbb{Y}$.
  - Analizar residuales para validar el modelo.
  - Acompañar la estimación de β con una región de confianza que involucra $(\mathbb{X}^T\mathbb{X})^{-1}$.
  - Predecir para un nuevo $\mathbf{X}_N$: $\hat{Y}_N = \hat{\beta}^T\mathbf{X}_N$.

Regresión lineal. $n > p$ y p fijo

- Modelo teórico: $Y|\mathbf{X} = \beta^T\mathbf{X} + \epsilon$, $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^p$, p fijo, $\epsilon \sim N(0, \sigma^2)$.
- ¿β? En población $\beta = \Sigma^{-1}\Sigma_{XY}$ con $\Sigma = \mathrm{var}(\mathbf{X})$ y $\Sigma_{XY} = \mathrm{cov}(\mathbf{X}, Y)$.
- Dado un conjunto de datos $(\mathbf{X}_i, Y_i) \in \mathbb{R}^{p+1}$, $i = 1, \ldots, n$, $\mathbb{X} \in \mathbb{R}^{n\times p}$, $\mathbb{Y} \in \mathbb{R}^{n\times 1}$ siguiendo el modelo:
  - Estimar β: $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$, con $\widehat{\Sigma} = \frac{1}{n}\mathbb{X}^T\mathbb{X}$ y $\widehat{\Sigma}_{XY} = \frac{1}{n}\mathbb{X}^T\mathbb{Y}$.
  - Analizar residuales para validar el modelo.
  - Acompañar la estimación de β con una región de confianza que involucra $(\mathbb{X}^T\mathbb{X})^{-1}$.
  - Predecir para un nuevo $\mathbf{X}_N$: $\hat{Y}_N = \hat{\beta}^T\mathbf{X}_N$.
  - ¿$\hat{\beta}$ y $\hat{\beta}^T\mathbf{X}_N$ son consistentes? (cuando n crece).

¿Qué pasa cuando p crece? ¿Por qué crece?

- Supongamos que cada vez tenemos más información del sujeto: $\mathbb{X}$ gana columnas.

¿Qué pasa cuando p crece? ¿Por qué crece?

- Supongamos que cada vez tenemos más información del sujeto: $\mathbb{X}$ gana columnas.
- Esto debería ayudar a estimar. Pero si la cantidad de sujetos no crece al mismo ritmo, llega un punto en que no podemos invertir $\mathbb{X}^T\mathbb{X}$. (Recordar $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}\mathbb{Y}$.)

¿Qué pasa cuando p crece? ¿Por qué crece?

- Supongamos que cada vez tenemos más información del sujeto: $\mathbb{X}$ gana columnas.
- Esto debería ayudar a estimar. Pero si la cantidad de sujetos no crece al mismo ritmo, llega un punto en que no podemos invertir $\mathbb{X}^T\mathbb{X}$. (Recordar $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}\mathbb{Y}$.)
- Aun con $n > p$, si $p \sim n$, $\mathbb{X}^T\mathbb{X}$ es casi singular y, por ende, la varianza del estimador de mínimos cuadrados (orden de $(\mathbb{X}^T\mathbb{X})^{-1}$) es tan grande que la estimación es más una incertidumbre que una certeza. Sin embargo...

Objetivos

- Nuestro objetivo puede ser: **estimación** (estimar β) o **predicción** (predecir Y para un nuevo $\mathbf{X}_N$). Están relacionados (obvio).
- En **estimación**, ¿cómo lograr consistencia cuando n y p crecen?, supongamos $n > p$ (¿por qué?).

Objetivos

- Nuestro objetivo puede ser: **estimación** (estimar β) o **predicción** (predecir Y para un nuevo $\mathbf{X}_N$). Están relacionados (obvio).
- En **estimación**, ¿cómo lograr consistencia cuando n y p crecen?, supongamos $n > p$ (¿por qué?).
  - $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$,

Objetivos

- Nuestro objetivo puede ser: **estimación** (estimar β) o **predicción** (predecir Y para un nuevo $\mathbf{X}_N$). Están relacionados (obvio).
- En **estimación**, ¿cómo lograr consistencia cuando n y p crecen?, supongamos $n > p$ (¿por qué?).
  - $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$,
  - la consistencia para p fijo es consecuencia de la consistencia de $\widehat{\Sigma}^{-1}$ y $\widehat{\Sigma}_{XY}$.

Objetivos

- Nuestro objetivo puede ser: **estimación** (estimar β) o **predicción** (predecir Y para un nuevo $\mathbf{X}_N$). Están relacionados (obvio).
- En **estimación**, ¿cómo lograr consistencia cuando n y p crecen?, supongamos $n > p$ (¿por qué?).
  - $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$,
  - la consistencia para p fijo es consecuencia de la consistencia de $\widehat{\Sigma}^{-1}$ y $\widehat{\Sigma}_{XY}$.
  - Johnstone y Lu (2009) probaron que $\widehat{\Sigma} \to \Sigma$ con error de orden p/n. Consistencia si $p/n \to 0$. Lo mismo para $\|\widehat{\Sigma}_{XY} - \Sigma_{XY}\|$.

Objetivos

- Nuestro objetivo puede ser: **estimación** (estimar β) o **predicción** (predecir Y para un nuevo $\mathbf{X}_N$). Están relacionados (obvio).
- En **estimación**, ¿cómo lograr consistencia cuando n y p crecen?, supongamos $n > p$ (¿por qué?).
  - $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$,
  - la consistencia para p fijo es consecuencia de la consistencia de $\widehat{\Sigma}^{-1}$ y $\widehat{\Sigma}_{XY}$.
  - Johnstone y Lu (2009) probaron que $\widehat{\Sigma} \to \Sigma$ con error de orden p/n. Consistencia si $p/n \to 0$. Lo mismo para $\|\widehat{\Sigma}_{XY} - \Sigma_{XY}\|$.
  - Sin embargo, esto no garantiza la (no) consistencia del producto...

Predicción. Qué buscamos

- Modelo

$$Y = \beta_{p_0}^T \mathbf{X}_{p_0} + b_{p_0+1} X_{p_0+1} + b_{p_0+2} X_{p_0+2} + \cdots + b_p X_p + \epsilon, \quad \epsilon \sim N(0, \sigma_p^2)$$

- Objetivo: consistencia de la predicción. Dado un nuevo $\mathbf{X}_N$,

¿$\hat{Y}_N = \hat{\beta}^T \mathbf{X}_N$ está cerca del *verdadero* $\beta^T \mathbf{X}_N$?

Una simulación

Tres ejemplos para trabajar: $(Y, \mathbf{X}_p)$ multivariados (la distribución de Y es fija; agregamos predictores sin cambiar la distribución de los previos) y tomamos $n = 2p > p$, $p = 2^4, 2^5, \ldots, 2^{10}$.

- Escenario 1: incorporamos X_i que no aportan información sobre Y.

Una simulación

Tres ejemplos para trabajar: $(Y, \mathbf{X}_p)$ multivariados (la distribución de Y es fija; agregamos predictores sin cambiar la distribución de los previos) y tomamos $n = 2p > p$, $p = 2^4, 2^5, \ldots, 2^{10}$.

- Escenario 1: incorporamos X_i que no aportan información sobre Y.
- Escenario 2: incorporamos X_i que agregan algo de información.

Una simulación

Tres ejemplos para trabajar: $(Y, \mathbf{X}_p)$ multivariados (la distribución de Y es fija; agregamos predictores sin cambiar la distribución de los previos) y tomamos $n = 2p > p$, $p = 2^4, 2^5, \ldots, 2^{10}$.

- Escenario 1: incorporamos X_i que no aportan información sobre Y.
- Escenario 2: incorporamos X_i que agregan algo de información.
- Escenario 3: incorporamos X_i que acumulan cada vez más información.

Una simulación

Tres ejemplos para trabajar: $(Y, \mathbf{X}_p)$ multivariados (la distribución de Y es fija; agregamos predictores sin cambiar la distribución de los previos) y tomamos $n = 2p > p$, $p = 2^4, 2^5, \ldots, 2^{10}$.

- Escenario 1: incorporamos X_i que no aportan información sobre Y.
- Escenario 2: incorporamos X_i que agregan algo de información.
- Escenario 3: incorporamos X_i que acumulan cada vez más información.

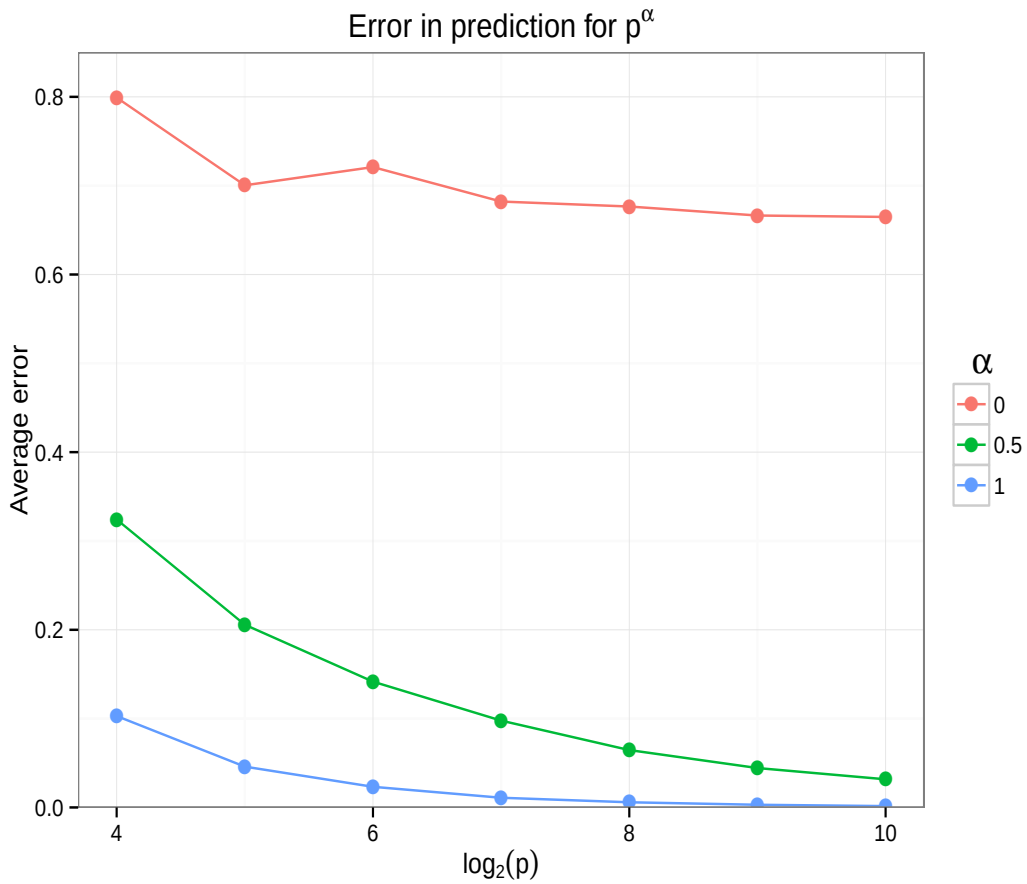
Aproximadamente p^α con $\alpha = 0, 0.5, 1$ predictores son informativos para Y. $\alpha = 0$ en el 1, $\alpha = .5$ en el 2 y $\alpha = 1$ en el 3.

Simulación. Más

Para esos p y n repetimos (muchas veces) la generación de muestras con la misma distribución (para cada par (p, n)). Estimamos β por mínimos cuadrados.

Predicción: estudiamos $|\hat{\beta}^T \mathbf{X}_N - \beta^T \mathbf{X}_N|$ para una nueva muestra $\mathbf{X}_N$ y reportamos el error cuadrático medio.

Resultados del experimento. $n = 2p$



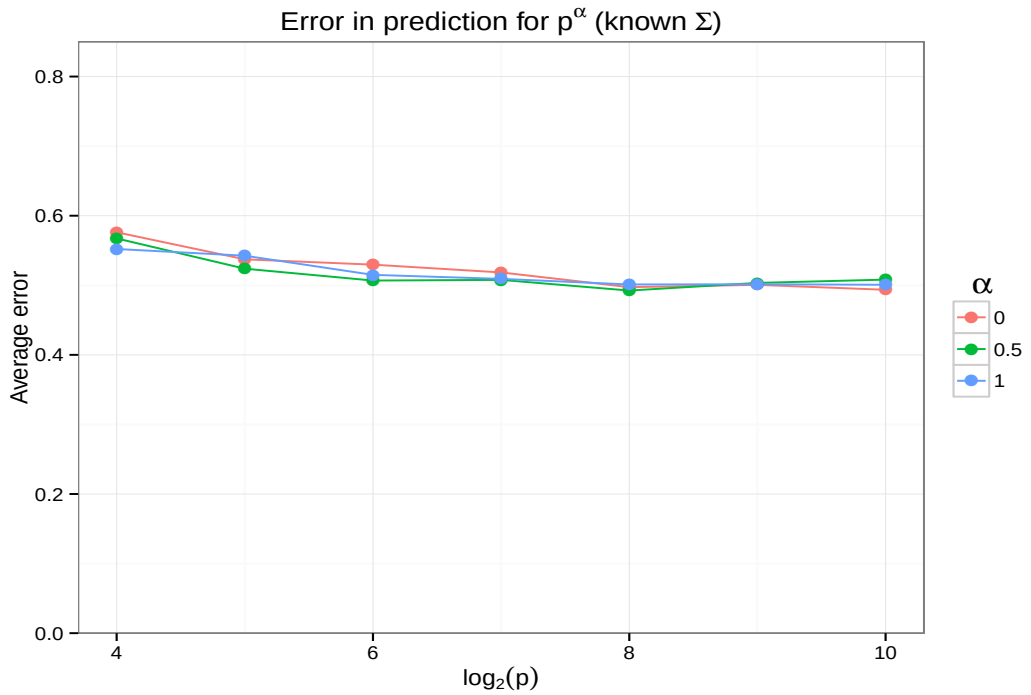
Una curiosidad. ¿Qué pasa si conocemos $\Sigma = \text{var}(\mathbf{X}_p)$?

¿Qué ocurre si conocemos la Σ verdadera y usamos

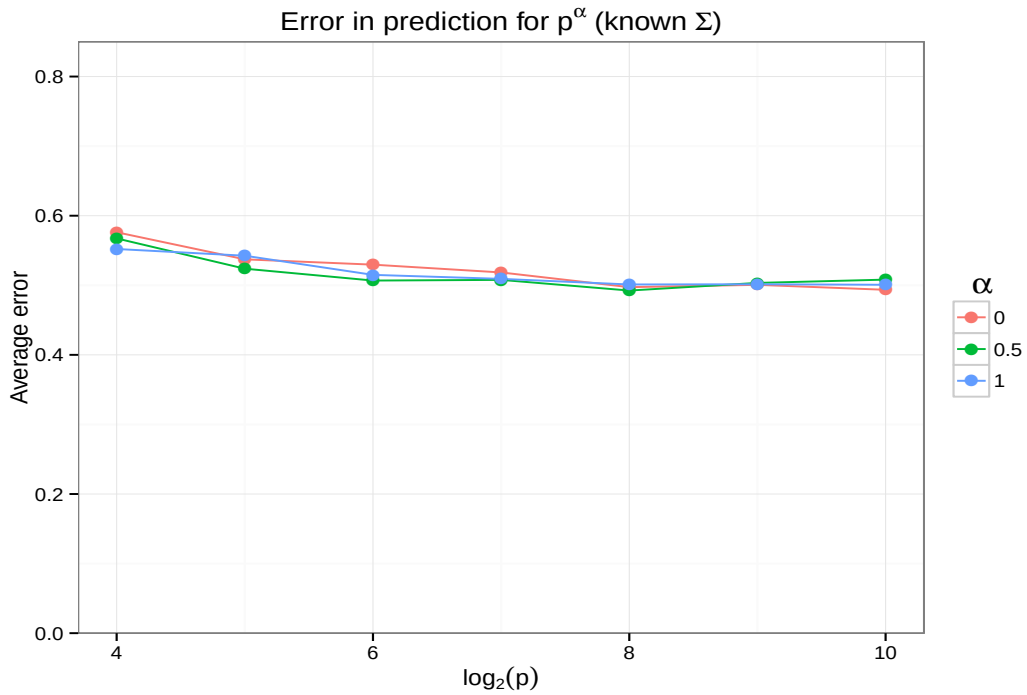
$$\hat{\beta} = \Sigma^{-1}\widehat{\Sigma}_{XY} \ ?$$

¿Da una mejor respuesta?

Resultados asumiendo varianza conocida. $n > p$



Resultados asumiendo varianza conocida. $n > p$



Moraleja: cuando es un parámetro "molesto", ¡estimá Σ aunque la sepas!

Predicción. Cuándo funciona

- Escenario 1: en realidad, pocos predictores tienen información sobre Y. Se puede *cambiar* a estimadores pensados para *esparsidad* (lasso, LAR, lasso adaptativo, elastic net).

- Escenarios 2 y 3: los métodos *esparsos* fallan; la regresión es lo opuesto: **abundante**. ¿Cómo definir "regresión abundante"?

Regresión esparsa

"Esparsa" significa que, aunque agregues predictores, sólo unos pocos quedan activos en la regresión. Es decir, al agregar muchos, agregás ruido.

Consistencia de estimadores y predicción con lasso, etc.: bajo ciertas restricciones.

¿Por qué esparsidad?

Hay contextos donde la esparsidad viene de la ciencia subyacente; algunos la ven como ley natural: si la regresión es de alta dimensión, "debe" ser esparsa.

Otros la vieron como único recurso (principio "bet-on-sparsity" de Bartlett et al., 2004): no sabemos estimar en el otro caso.

Bajo esparsidad, la selección de variables evita acumulación de ruido, mejora predicción y hace el modelo más interpretable.

Y en el mismo año, en otra comunidad

Esparsidad vs Abundancia. Cita de *Hierarchical multiblock PLS and PC models...* (Wold, Kettaneh, Tjessem, 1996). ¿Quién es Wold?

Y en el mismo año, en otra comunidad

Esparsidad vs Abundancia. Cita de *Hierarchical multiblock PLS and PC models...* (Wold, Kettaneh, Tjessem, 1996). ¿Quién es Wold?

> **En situaciones con muchas variables (50–100+), hay una fuerte tentación de reducir drásticamente su número... Sin embargo, esa reducción suele quitar información, sesgar la interpretación y aumentar el riesgo de modelos espurios. Una alternativa mejor que eliminar variables es** *dividirlas en bloques con sentido conceptual y aplicar modelos PLS/PC multibloque jerárquicos...*

Y en el mismo año, en otra comunidad

Esparsidad vs Abundancia. Cita de *Hierarchical multiblock PLS and PC models...* (Wold, Kettaneh, Tjessem, 1996). ¿Quién es Wold?

> **En situaciones con muchas variables (50–100+), hay una fuerte tentación de reducir drásticamente su número... Sin embargo, esa reducción suele quitar información, sesgar la interpretación y aumentar el riesgo de modelos espurios. Una alternativa mejor que eliminar variables es** *dividirlas en bloques con sentido conceptual y aplicar modelos PLS/PC multibloque jerárquicos...*
>
> *Con PLS y PCA la situación es distinta: funcionan bien aun con muchas variables y N pequeño. De hecho, cuantas más variables relevantes, más precisas las puntuaciones t (y u en PLS), pues son promedios ponderados y los promedios mejoran con más elementos.* **No hay necesidad real de mantener pocas variables; sólo deben eliminarse las realmente irrelevantes.**

¿De qué habla Wold? Definición de *regresión abundante*

Una regresión es abundante si $R^2_{Y\mathbf{X}_p} \to 1$ *cuando* $p \to \infty$, *donde* $R_{Y\mathbf{X}_p}$ *es el coeficiente de correlación múltiple entre* $\mathbf{X}_p$ *y* Y *(la contribución de* $\mathbf{X}_p$ *en* Y *crece —aunque sea un poco— con* p*).*

Coeficiente de abundancia

Definimos el *coeficiente de abundancia*:

$$h(p) = \frac{R^2_{Y\mathbf{x}_p}}{1 - R^2_{Y\mathbf{x}_p}}.$$

Coeficiente de abundancia

Definimos el *coeficiente de abundancia*:

$$h(p) = \frac{R^2_{Y\mathbf{x}_p}}{1 - R^2_{Y\mathbf{x}_p}}.$$

$h(p) \sim 1$ cuando hay esparsidad.
$h(p)$ crece cuando hay abundancia.

Teoría

$$\text{Sea } h(p) = \frac{R^2_{Y\mathbf{x}_p}}{1-R^2_{Y\mathbf{x}_p}} \text{ y } \mathbf{V} = E\left(\hat{\beta}^T(\mathbf{X}_N - \bar{X}) - \beta^T(\mathbf{X}_N - \mu_X)\right)^2.$$

Teoría

Sea $h(p) = \frac{R^2_{Y\mathbf{x}_p}}{1-R^2_{Y\mathbf{x}_p}}$ y $\mathbf{V} = E\left(\hat{\beta}^T(\mathbf{X}_N - \bar{X}) - \beta^T(\mathbf{X}_N - \mu_X)\right)^2$.

- Si $\hat{\beta} = \Sigma^{-1}\widehat{\Sigma}_{\mathbf{X}_p Y}$ entonces $\mathbf{V} = O_p(p/n)$.

Teoría

Sea $h(p) = \frac{R^2_{Y\mathbf{x}_p}}{1-R^2_{Y\mathbf{x}_p}}$ y $\mathbf{V} = E\left(\hat{\beta}^T(\mathbf{X}_N - \bar{X}) - \beta^T(\mathbf{X}_N - \mu_X)\right)^2$.

- Si $\hat{\beta} = \Sigma^{-1}\widehat{\Sigma}_{\mathbf{x}_p Y}$ entonces $\mathbf{V} = O_p(p/n)$.
- Si $n > p+2$ y $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$ entonces $\mathbf{V} = O_p\left(\frac{pn}{n\,h(p)\,(n-p-2)}\right)$ (simulaciones con $n \sim 2p$):
  - $h(p) \sim 1$ (esparsa) $\Rightarrow \mathbf{V} = O_p(p/n)$ (escenario 1).

Teoría

Sea $h(p) = \frac{R^2_{Y\mathbf{x}_p}}{1-R^2_{Y\mathbf{x}_p}}$ y $\mathbf{V} = E\left(\hat{\beta}^T(\mathbf{X}_N - \bar{X}) - \beta^T(\mathbf{X}_N - \mu_X)\right)^2$.

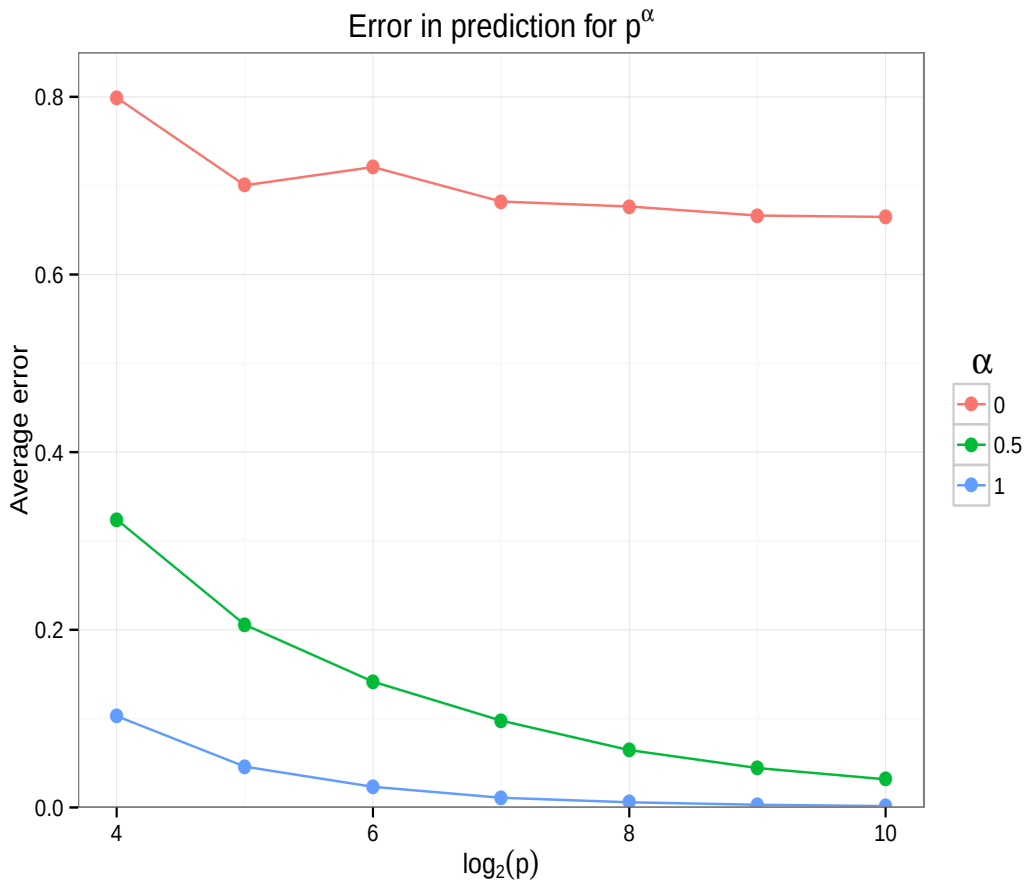
- Si $\hat{\beta} = \Sigma^{-1}\widehat{\Sigma}_{\mathbf{X}_p Y}$ entonces $\mathbf{V} = O_p(p/n)$.
- Si $n > p+2$ y $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$ entonces $\mathbf{V} = O_p\left(\frac{pn}{n\,h(p)\,(n-p-2)}\right)$ (simulaciones con $n \sim 2p$):
  - $h(p) \sim 1$ (esparsa) $\Rightarrow \mathbf{V} = O_p(p/n)$ (escenario 1).
  - $h(p) \sim p \Rightarrow \mathbf{V} = O_p(1/n)$ (escenario 3).

Teoría

Sea $h(p) = \frac{R^2_{Y\mathbf{x}_p}}{1-R^2_{Y\mathbf{x}_p}}$ y $\mathbf{V} = E\left(\hat{\beta}^T(\mathbf{X}_N - \bar{X}) - \beta^T(\mathbf{X}_N - \mu_X)\right)^2$.

- Si $\hat{\beta} = \Sigma^{-1}\widehat{\Sigma}_{\mathbf{x}_p Y}$ entonces $\mathbf{V} = O_p(p/n)$.
- Si $n > p+2$ y $\hat{\beta} = \widehat{\Sigma}^{-1}\widehat{\Sigma}_{XY}$ entonces $\mathbf{V} = O_p\left(\frac{pn}{n\,h(p)\,(n-p-2)}\right)$ (simulaciones con $n \sim 2p$):
  - $h(p) \sim 1$ (esparsa) $\Rightarrow \mathbf{V} = O_p(p/n)$ (escenario 1).
  - $h(p) \sim p \Rightarrow \mathbf{V} = O_p(1/n)$ (escenario 3).
  - $h(p) \sim p^\alpha$ con $\alpha < 1 \Rightarrow \mathbf{V} = O_p(1/n^\alpha)$ (escenario 2).

Resultados del experimento. $n = 2p$



¿Qué pasa si $p > n$?

- Recordemos: para $p < n$, $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}\mathbb{Y}$.
- Si $p > n$, $\mathbb{X}^T\mathbb{X}$ no es invertible.
- Un enfoque: estimadores *esparsos*. Pero ¿y si la regresión no es esparsa?
- Podemos tomar *una* inversa generalizada y definir

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-}\mathbb{X}^T\mathbb{Y}.$$

¿Qué pasa si $p > n$?

- Recordemos: para $p < n$, $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}\mathbb{Y}$.
- Si $p > n$, $\mathbb{X}^T\mathbb{X}$ no es invertible.
- Un enfoque: estimadores *esparsos*. Pero ¿y si la regresión no es esparsa?
- Podemos tomar *una* inversa generalizada y definir

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-}\mathbb{X}^T\mathbb{Y}.$$

- Predicción *in-sample*: para *cualquier* inversa generalizada, $\hat{Y}$ no cambia (aunque sí $\hat{\beta}$).
- Predicción *out-of-sample*: no es así.

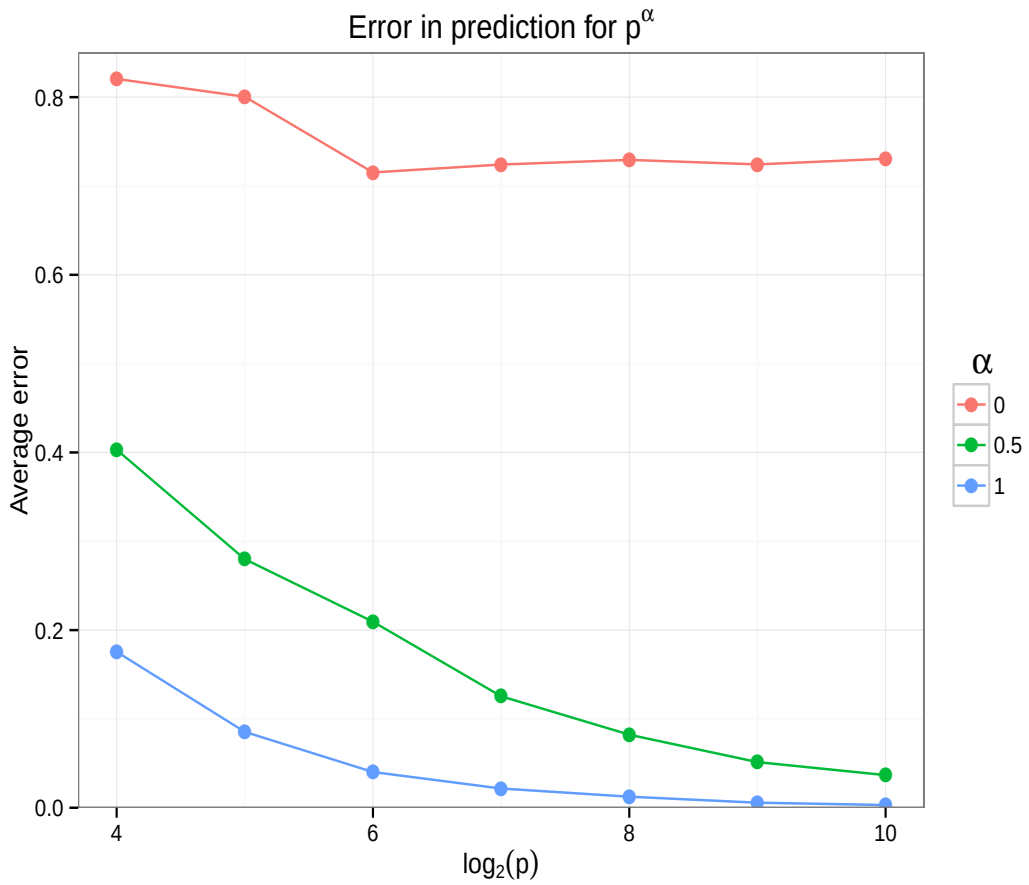
¿Qué pasa si $p > n$?

- Recordemos: para $p < n$, $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}\mathbb{Y}$.
- Si $p > n$, $\mathbb{X}^T\mathbb{X}$ no es invertible.
- Un enfoque: estimadores *esparsos*. Pero ¿y si la regresión no es esparsa?
- Podemos tomar *una* inversa generalizada y definir

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-}\mathbb{X}^T\mathbb{Y}.$$

- Predicción *in-sample*: para *cualquier* inversa generalizada, $\hat{Y}$ no cambia (aunque sí $\hat{\beta}$).
- Predicción *out-of-sample*: no es así.
- ¿Qué inversa usar? ¿Qué podemos calcular?

Resultados con inversa generalizada y abundancia. $p = 2n$



Sobre los resultados

Las herramientas para probar resultados suelen asumir autovalores acotados de Σ.

Sobre los resultados

Las herramientas para probar resultados suelen asumir autovalores acotados de Σ.
¿Podemos tener regresión abundante y, a la vez, Σ con autovalores acotados?

$$\Sigma = E(\mathrm{cov}(\mathbf{X}_p|Y)) + \Sigma_{X_pY}\Sigma_{X_pY}^T/\sigma_Y^2.$$

Sobre los resultados

Las herramientas para probar resultados suelen asumir autovalores acotados de Σ.
¿Podemos tener regresión abundante y, a la vez, Σ con autovalores acotados?

$$\Sigma = E(\mathrm{cov}(\mathbf{X}_p|Y)) + \Sigma_{X_pY}\Sigma_{X_pY}^T/\sigma_Y^2.$$

Autovalores acotados de Σ implican Σ_{X_pY} acotada, lo que choca con $R_{X_pY} \to 1$ (abundancia).

Sobre los resultados

¿Abundancia y a la vez Σ con autovalores acotados?

$$\Sigma = E(\mathrm{cov}(\mathbf{X}_p|Y)) + \Sigma_{X_pY}\Sigma_{X_pY}^T/\sigma_Y^2.$$

Esto implicaría Σ_{X_pY} acotada, contradictorio con $R_{X_pY} \to 1$.

Parece importante que Σ tenga autovalores no acotados para lograr consistencia en predicción.

Para la prueba se necesita $\mathrm{var}((\mathbb{X}^T\mathbb{X})^-)$, problema abierto cuando la varianza verdadera de $\mathbf{X}$ no es $\sigma^2 I_p$.

Resumiendo

Para $n < p$ no hay tanto progreso:

- Resultados negativos si Σ tiene autovalores acotados usando la inversa de Penrose o Σ conocida —pero ahí no hay **abundancia**.
- Sin resultados positivos probados cuando Σ tiene autovalores no acotados (caso **abundante**); las simulaciones se ven muy bien.
- ¿Próximos pasos?

Mirada de regresión inversa

¿Abundancia y a la vez Σ con autovalores acotados? No, pero si miramos

$$\Sigma = E(\mathrm{cov}(\mathbf{X}_p|Y)) + \Sigma_{X_pY}\Sigma_{X_pY}^T/\sigma_Y^2$$

y pedimos que $\Delta := E(\mathrm{cov}(\mathbf{X}_p|Y))$ tenga autovalores acotados, podemos avanzar.

Como $\beta = \Sigma^{-1}\Sigma_{XY}$,

Mirada de regresión inversa

¿Abundancia y a la vez Σ con autovalores acotados? No, pero si miramos

$$\Sigma = E(\mathrm{cov}(\mathbf{X}_p|Y)) + \Sigma_{X_pY}\Sigma_{X_pY}^T/\sigma_Y^2$$

y pedimos que $\Delta := E(\mathrm{cov}(\mathbf{X}_p|Y))$ tenga autovalores acotados, podemos avanzar.

Como $\beta = \Sigma^{-1}\Sigma_{XY}$,

$$\beta = \Delta^{-1}\Sigma_{XY}/\left(1 + \Sigma_{XY}^T\Delta^{-1}\Sigma_{XY}\right).$$

Resultados de SDR (regresión inversa)

- Si Δ tiene autovalores acotados, estimamos $\widehat{\Delta}$ mediante *alguna* inversa generalizada o estimadores de covarianza para $n < p$ (hay cientos).
- Probamos consistencia para $\hat{\beta}^T \mathbf{X}_N$ (no para $\hat{\beta}$), con

$$\hat{\beta} = \widehat{\Delta}^{-1}\hat{\Sigma}_{XY} / \left(1 + \hat{\Sigma}_{XY}^T \widehat{\Delta}^{-1}\hat{\Sigma}_{XY}\right)$$

cuando hay abundancia. Como

$$\Sigma = \Delta + \Sigma_{X_p Y}\Sigma_{X_p Y}^T/\sigma_Y^2,$$

abundancia significa $\|\Sigma_{X_p Y}\| \to \infty$.

Más sobre abundancia. Regresión PLS

- PLS es de los primeros métodos de predicción en regresiones lineales de alta dimensión (n no grande respecto de p).

Más sobre abundancia. Regresión PLS

- PLS es de los primeros métodos de predicción en regresiones lineales de alta dimensión (n no grande respecto de p).
- Iniciado por Herman Wold (años 60) y adaptado por Svante Wold (1977) para quimiometría.

Más sobre abundancia. Regresión PLS

- PLS es de los primeros métodos de predicción en regresiones lineales de alta dimensión (n no grande respecto de p).
- Iniciado por Herman Wold (años 60) y adaptado por Svante Wold (1977) para quimiometría.
- En quimiometría, donde la *predicción* es central, PLS es método de cabecera.

Más sobre abundancia. Regresión PLS

- PLS es de los primeros métodos de predicción en regresiones lineales de alta dimensión (n no grande respecto de p).
- Iniciado por Herman Wold (años 60) y adaptado por Svante Wold (1977) para quimiometría.
- En quimiometría, donde la *predicción* es central, PLS es método de cabecera.
- Suelen no plantear modelos poblacionales ni coeficientes, sino trabajar directo con algoritmos de predicción.

Más sobre PLS

- Como algoritmo para predecir en $n < p$ o $n \sim p$, sin modelo explícito, las asintóticas y otros constructos estadísticos tardaron en aparecer.

Más sobre PLS

- Como algoritmo para predecir en $n < p$ o $n \sim p$, sin modelo explícito, las asintóticas y otros constructos estadísticos tardaron en aparecer.
- Aun sin "teoría detrás", es central en quimiometría.

¿PLS funciona? Martens & Næs (1989)

PLS surgió para evitar (cuando $n < p$) invertir Σ en $\beta = \Sigma^{-1}\Sigma_{XY}$ del modelo $Y = \beta^T \mathbf{X} + \epsilon$.

¿PLS funciona? Martens & Næs (1989)

PLS surgió para evitar (cuando $n < p$) invertir Σ en $\beta = \Sigma^{-1}\Sigma_{XY}$ del modelo $Y = \beta^T \mathbf{X} + \epsilon$.
Versión simplificada del algoritmo:

- Elegir d (hay formas de elegirlo).

¿PLS funciona? Martens & Næs (1989)

PLS surgió para evitar (cuando $n < p$) invertir Σ en $\beta = \Sigma^{-1}\Sigma_{XY}$ del modelo $Y = \beta^T \mathbf{X} + \epsilon$.
Versión simplificada del algoritmo:

- Elegir d (hay formas de elegirlo).
- Calcular $\hat{S} = \{\hat{\Sigma}_{\mathbf{X}Y}, \ldots, \hat{\Sigma}^{d-1}\hat{\Sigma}_{\mathbf{X}Y}\}$ con versiones muestrales.

¿PLS funciona? Martens & Næs (1989)

PLS surgió para evitar (cuando $n < p$) invertir Σ en $\beta = \Sigma^{-1}\Sigma_{XY}$ del modelo $Y = \beta^T \mathbf{X} + \epsilon$.
Versión simplificada del algoritmo:

- Elegir d (hay formas de elegirlo).
- Calcular $\hat{S} = \{\hat{\Sigma}_{\mathbf{X}Y}, \ldots, \hat{\Sigma}^{d-1}\hat{\Sigma}_{\mathbf{X}Y}\}$ con versiones muestrales.
- Elegir $\hat{\beta} \in \text{span}(\hat{S})$ que minimice $\|\mathbb{Y} - \mathbb{X}\hat{\beta}\|$.

Se puede probar (Helland) que $\hat{\beta} = \hat{S}(\hat{S}^T\hat{\Sigma}\hat{S})^{-1}\hat{S}^T\hat{\Sigma}_{\mathbf{X}Y}$.

A posteriori podemos ver que es una forma aproximada de resolver $\Sigma\beta = \Sigma_{XY}$.

PLS funciona

- Funciona incluso con $n < p$, pero faltaba teoría que explicara por qué.

PLS funciona

- Funciona incluso con $n < p$, pero faltaba teoría que explicara por qué.
- La comunidad estadística prestó poca atención al principio (quizás por la falta de modelo explícito).

Luego sí aparecieron los estadísticos

Resultados positivos (Cook, Helland y Su):

- En población, para $d = 1$: $\beta = \Sigma_{\mathbf{X}Y}(\Sigma_{\mathbf{X}Y}^T \Sigma \Sigma_{\mathbf{X}Y})^{-1}\Sigma_{\mathbf{X}Y}^T \Sigma_{XY}$, lo que implica:
  - $\beta = c\,\Sigma_{\mathbf{X}Y}$,
  - $\Sigma_{\mathbf{X}Y}$ es autovector de Σ. ¿ Por qué? $\Sigma\beta = \Sigma_{XY}$ y $\beta = c\Sigma_{XY}$.

Luego sí aparecieron los estadísticos

Resultados positivos (Cook, Helland y Su):

- En población, para $d = 1$: $\beta = \Sigma_{\mathbf{X}Y}(\Sigma_{\mathbf{X}Y}^T \Sigma \Sigma_{\mathbf{X}Y})^{-1}\Sigma_{\mathbf{X}Y}^T \Sigma_{XY}$, lo que implica:
  - $\beta = c\,\Sigma_{\mathbf{X}Y}$,
  - $\Sigma_{\mathbf{X}Y}$ es autovector de Σ. ¿ Por qué? $\Sigma\beta = \Sigma_{XY}$ y $\beta = c\Sigma_{XY}$.

  Además, si $d > 1$, β "corta" sólo d autovectores de Σ, es decir, β vive en la envolvente generada por d autovectores.

Modelo

Si $Y = \beta^T \mathbf{X} + \varepsilon$ y existe $\Gamma \in \mathbb{R}^{p \times d}$ tal que

- $\beta = \Gamma A$ (para algún A),
- $\Sigma = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T$ (siendo Γ_0 complemento ortogonal),

entonces, con p fijo y $n \to \infty$, PLS es consistente para β y, por ende, la predicción es consistente. (El modelo siempre es cierto al menos con $d = p$.)

Pero en quimiometría p crece

- ¿Y si p crece? El algoritmo funciona si $d < \min\{p, n\}$, incluso cuando $n < p$.

Pero en quimiometría p crece

- ¿Y si p crece? El algoritmo funciona si $d < \min\{p, n\}$, incluso cuando $n < p$.
- A la vista del éxito práctico, cabe esperar buenas propiedades estadísticas en alta dimensión.

Aparecen de nuevo... con malas noticias

Chun & Keleş mostraron que, en cierto marco, el estimador PLS es inconsistente salvo que $p/n \to 0$; motivaron versiones *esparsas* de PLS.

Un dilema

- Décadas de uso avalan a PLS, pero su inconsistencia con $p/n \to c > 0$ tensiona su uso en alta dimensión.
- Posibles explicaciones:
  - la consistencia no siempre predice el valor práctico;
  - la literatura sobre PLS podría sobrestimar su valor;
  - el constructo de Chun–Keleş no refleja el rango real de aplicaciones.

Modelo en Chun-Keleş

$$X|Y = \mu_X + \Theta\nu_y + \omega,$$

donde $\nu \in \mathbb{R}^d$, $\nu \sim N(0, I_d)$, $\Theta \in \mathbb{R}^{p\times d}$, $\omega \in \mathbb{R}^p$, y (ruido) $w \sim N(0, \pi^2 I_p)$.

Modelo en Chun-Keleş

$$X|Y = \mu_X + \Theta\nu_y + \omega,$$

donde $\nu \in \mathbb{R}^d$, $\nu \sim N(0, I_d)$, $\Theta \in \mathbb{R}^{p\times d}$, $\omega \in \mathbb{R}^p$, y (ruido) $w \sim N(0, \pi^2 I_p)$.
Entonces $X \perp\!\!\!\perp Y \mid \Theta^T X$; d combinaciones lineales llevan toda la info de X sobre Y.

$$\Sigma = \Theta\Theta^T + \pi^2 I_p = H(\Theta^T\Theta + \pi^2 I_d)H^T + \pi^2 Q_H.$$

Supuestos en Chun-Keleş

$$\Sigma = \Theta\Theta^T + \pi^2 I_p = H(\Theta^T\Theta + \pi^2 I_d)H^T + \pi^2 Q_H.$$

- Columnas de Θ ortogonales con normas acotadas que convergen $\Rightarrow \Sigma$ acotada.
- En espectroscopía, es plausible que mucho "señal" venga de muchas longitudes de onda: muchas filas de Θ no nulas y $\sum_{i=1}^{p} \|\theta_i\|^2$ diverge. No se cumplen sus supuestos.

Conclusión: el paper fuerza *esparsidad* para obtener *no-consistencia*.

Vuelven... con buenas noticias

Bajo el mismo modelo,

$$X|Y = \mu_X + \Theta\nu_y + \omega, \quad \Sigma = \Theta\Theta^T + \pi^2 I_p,$$

la tasa del error cuadrático de predicción con PLS es

$$\frac{p}{\left(\sum_{i=1}^{p} \|\theta_i\|^2\right) n}.$$

Consecuencias

- Caso Chun–Keleş: $\sum \|\theta_i\|^2$ acotada $\Rightarrow$ consistencia sólo si $p/n \to 0$.

Consecuencias

- Caso Chun–Keleş: $\sum \|\theta_i\|^2$ acotada $\Rightarrow$ consistencia sólo si $p/n \to 0$.
- Si $\sum \|\theta_i\|^2 \sim p^\alpha$ (abundancia), el error $\sim \dfrac{p^{1-\alpha}}{n}$.

Cuando la info se acumula al máximo ($\sum \|\theta_i\|^2 \sim p$) hay consistencia tipo $\sqrt{n}$.

¿Más?

- Sí: resultados generales de consistencia (no sólo para Chun–Keleş).
- La tasa depende, grosso modo, de la razón entre la información nueva que aportan los predictores sobre Y y el ruido que agregan.

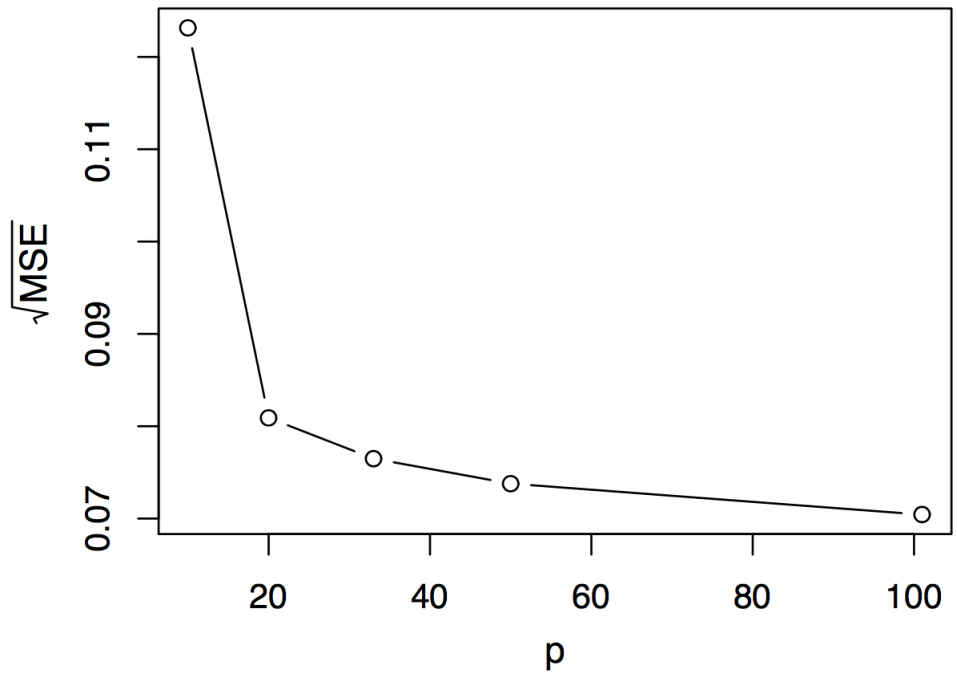
Datos de tetraciclina

- Goicoechea y Olivieri (1999) usan PLS para predecir concentración de tetraciclina en sangre humana. 50 muestras de entrenamiento (0–4 $\mu g\,mL^{-1}$) y 57 de validación.
- Predictores: intensidades de fluorescencia en $p = 101$ puntos (450–550 nm).
- Vía LOO, el mejor d fue 4 combinaciones lineales.

Tetraciclina: protocolo

- Ilustramos el comportamiento de PLS cuando p aumenta.
- PLS con $d = 4$ para predecir validación usando p espectros equiespaciados, p entre 10 y 101. Reportamos RMSE.

RMSE para distintos p



Caída pronunciada del RMSE para $p < 30$ y luego descenso lento pero sostenido. Al ser predicción real, el RMSE no va a 0 con p creciente, como en algunas simulaciones.

¡Gracias!

Referencias

1. Basa, J., Cook, R. D., Forzani, L., & Marcos, M. (2022). Asymptotic distribution of one-component PLS regression estimators in high dimensions. *Canadian Journal of Statistics*.
2. Cook, R. D., Forzani, L. (2021). PLS regression algorithms in the presence of nonlinearity. *Chemometrics and Intelligent Laboratory Systems*, 213, 104307.
3. Cook, R. D.; Forzani, L. (2020). Envelopes: A new chapter in PLS. *Journal of Chemometrics*, 34(10), e3287.
4. Cook, R. D., Forzani, L. (2019). Partial Least Squares Prediction in High-Dimensional Regression. *Annals of Statistics*, 47(2), 884–908.
5. Cook, R. D., Forzani, L. (2018). Big data and PLS prediction. *Canadian Journal of Statistics*, 46(1), 62–78.
6. Cook, R. D., Forzani, L., Rothman, A. J. (2015). Comentarios... *The American Statistician*, 69(3).
7. Rothman, A. J., Forzani, L. (2014). Properties of optimizations... *Electronic Journal of Statistics*, 8(2):2693–2700.
8. Cook, R. D., Forzani, L., Rothman, A. J. (2013). Prediction in abundant high-dimensional linear regression. *Electronic Journal of Statistics*, 7(1), 3059–3088.
9. Cook, R. D., Forzani, L., Rothman, A. J. (2012). Estimating sufficient reductions... *Annals of Statistics*, 40(1), 353–384.