Test de esfericidad parcial en alta dimensión

¿Cuántas componentes importan?

Antonella Gieco

En colaboración con Liliana Forzani y Carlos Tolmasky

Seminario IMAL- Noviembre de 2016

Resumen

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

1 Introducción
- Motivación

2 Modelos Spiked
- Johnstone (2001)
- Estadística en alta dimensión

3 Estimando la dimensión
- Enfoques previos
- Nuestro objetivo: test LRT
- "Information theoretic": estadístico penalizado

4 Conclusión

Motivación

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

1. Matrices de covarianza con sólo unos pocos autovalores importantes.

Motivación

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

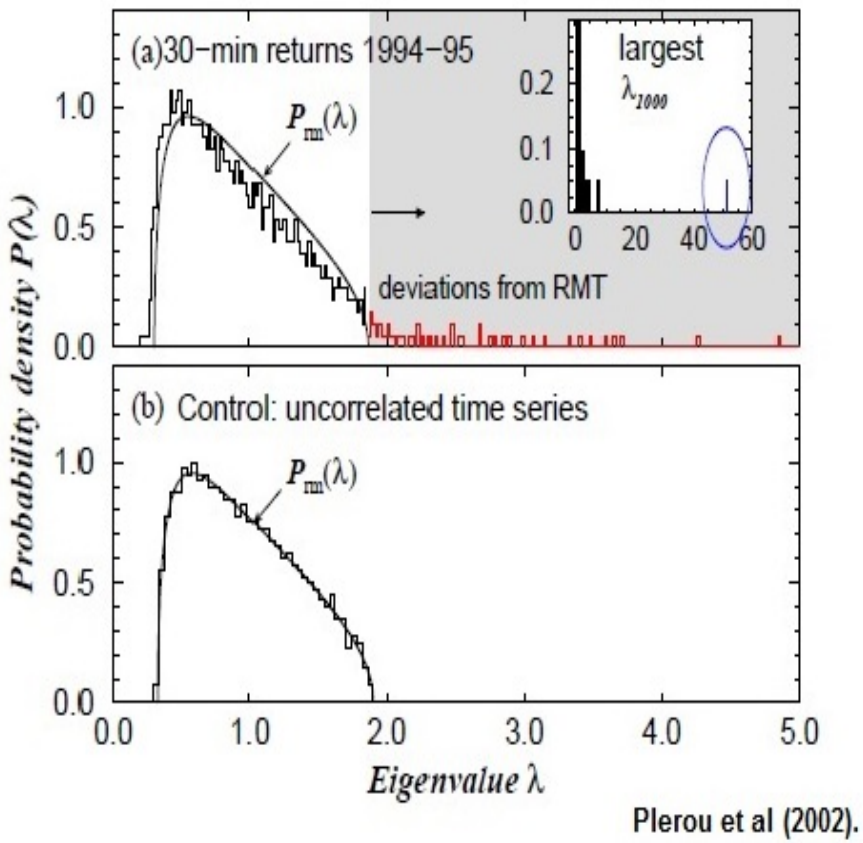
1. Matrices de covarianza con sólo unos pocos autovalores importantes.
2. El resto es ruido.

Motivación

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

1. Matrices de covarianza con sólo unos pocos autovalores importantes.
2. El resto es ruido.

- Ejemplo: stock market.
- El primer (mayor) autovalor es mucho más grande que el resto.
- Después de unos pocos primeros, el resto de ellos son pequeños.
- Plerou et al (2002).



Plerou et al (2002).

Modelos de covarianza Spiked (Johnstone, 2001)

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Este es un fenómeno común (no solo en finanzas).
- Uno o unos pocos ($d \ll p$) autovalores grandes y bien separados del resto
  $(\lambda_1 \geq \ldots \lambda_d \gg \lambda_{d+1} = \cdots = \lambda_p = \sigma^2)$.

$$\begin{pmatrix} \lambda_1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_d & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & \sigma^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & \ldots & \sigma^2 \end{pmatrix}$$

Alta dimensión vs Estadística clásica.

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Estadística multivariada "clásica":
  - Cantidad fija de variables (p) y muchas observaciones (n).
  - Dadas p variables, se pueden tener tantas observaciones como se quieran.
  - Resultados límites para $n \to \infty$.
- Estadística multivariada en "Alta-dimensión":
  - el número de variables (p) es comparable el número de observacions (n).
  - Resultados límites cuando $p/n \to y$ (una constante fija).

Algunos resultados en alta dimensión

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Distribución de los autovalores

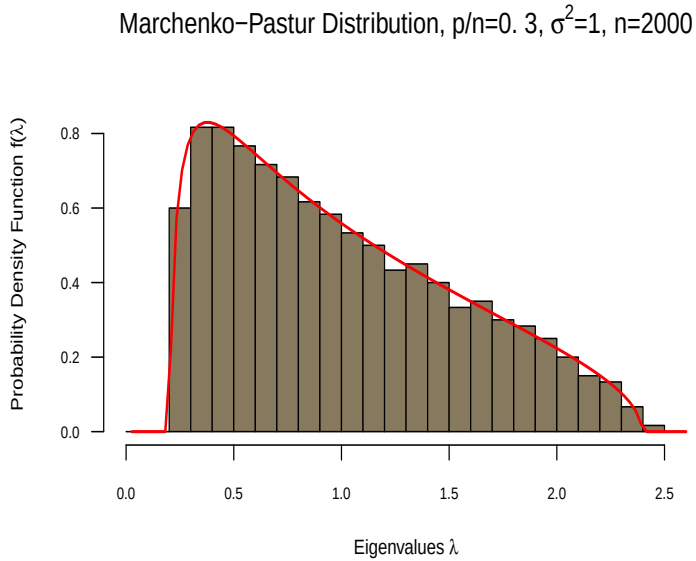
- p variables independientes con varianza σ^2
- n observaciones de cada una (p y n grandes).
- Los autovalores de la matriz de covarianzas muestral, no nos da cerca de $(\sigma^2, \sigma^2, \cdots, \sigma^2)$.

Algunos resultados en alta dimensión

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Distribución de los autovalores

- p variables independientes con varianza σ^2
- n observaciones de cada una (p y n grandes).
- Los autovalores de la matriz de covarianzas muestral, no nos da cerca de $(\sigma^2, \sigma^2, \cdots, \sigma^2)$.

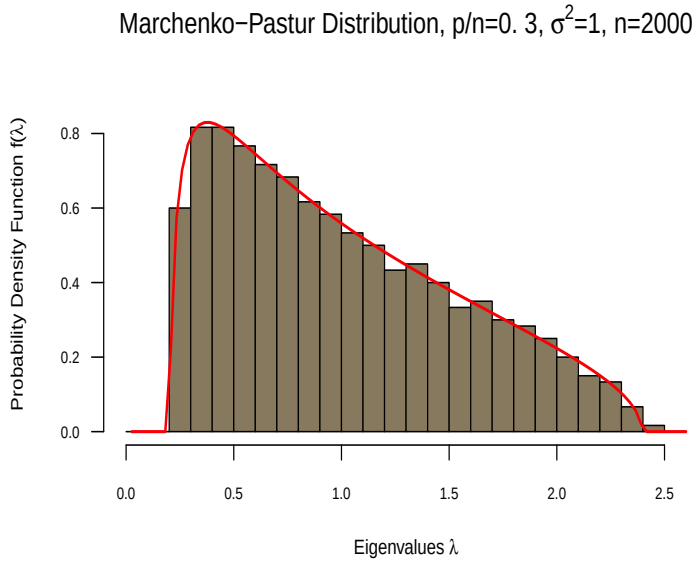


Algunos resultados en alta dimensión

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Distribución de los autovalores

- p variables independientes con varianza σ^2
- n observaciones de cada una (p y n grandes).
- Los autovalores de la matriz de covarianzas muestral, no nos da cerca de $(\sigma^2, \sigma^2, \cdots, \sigma^2)$.



La distribución límite (línea roja) es la M-P(y, σ^2).

El soporte es $\sigma^2 (1 \pm \sqrt{y})^2$ $(\approx (0.2\ ;\ 2.4))$.

Algunos resultados en alta dimensión

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Más resultados para estadística HD para modelos spiked.
  - Los estimadores de los λ's en el modelo de Johnstone tienen dos estados (Baik, J., G. B. Arous, and S. Péché (2005)):
    - $\hat{\lambda} \to \lambda\,(1 + \frac{y\sigma^2}{\lambda - \sigma^2})$ if $\lambda > \sigma^2\,(1 + \sqrt{y})$.
    - $\hat{\lambda} \to \sigma^2\,(1 + \sqrt{y})^2$ if $\lambda \in (\sigma^2, \sigma^2\,(1 + \sqrt{y})]$.

Algunos resultados en alta dimensión

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Más resultados para estadística HD para modelos spiked.
  - Los estimadores de los λ's en el modelo de Johnstone tienen dos estados (Baik, J., G. B. Arous, and S. Péché (2005)):
    - $\hat{\lambda} \to \lambda\,(1 + \frac{y\sigma^2}{\lambda - \sigma^2})$ if $\lambda > \sigma^2\,(1 + \sqrt{y})$.
    - $\hat{\lambda} \to \sigma^2\,(1 + \sqrt{y})^2$ if $\lambda \in (\sigma^2, \sigma^2\,(1 + \sqrt{y})]$.
  - Por lo tanto, si $\sigma^2 < \lambda \leq \sigma^2\,(1 + \sqrt{y})$ el estimador de λ no se separa del "montón".

Algunos resultados en alta dimensión

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Más resultados para estadística HD para modelos spiked.
  - Los estimadores de los λ's en el modelo de Johnstone tienen dos estados (Baik, J., G. B. Arous, and S. Péché (2005)):
    - $\hat{\lambda} \to \lambda\,(1 + \frac{y\sigma^2}{\lambda - \sigma^2})$ if $\lambda > \sigma^2\,(1 + \sqrt{y})$.
    - $\hat{\lambda} \to \sigma^2\,(1 + \sqrt{y})^2$ if $\lambda \in (\sigma^2, \sigma^2\,(1 + \sqrt{y})]$.
  - Por lo tanto, si $\sigma^2 < \lambda \leq \sigma^2\,(1 + \sqrt{y})$ el estimador de λ no se separa del "montón".
    Para alguna transformación ortogonal $\mathbf{U}$

$$\mathbf{U\Sigma U}^T = \text{diag}(\lambda_1, \lambda_2, \ldots, \ \ldots \lambda_d, \sigma^2, \ldots, \ \ldots, \sigma^2)$$

Algunos resultados en alta dimensión

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Más resultados para estadística HD para modelos spiked.
  - Los estimadores de los λ's en el modelo de Johnstone tienen dos estados (Baik, J., G. B. Arous, and S. Péché (2005)):
    - $\hat{\lambda} \to \lambda\,(1 + \frac{y\sigma^2}{\lambda - \sigma^2})$ if $\lambda > \sigma^2\,(1 + \sqrt{y})$.
    - $\hat{\lambda} \to \sigma^2\,(1 + \sqrt{y})^2$ if $\lambda \in (\sigma^2, \sigma^2\,(1 + \sqrt{y})]$.
  - Por lo tanto, si $\sigma^2 < \lambda \leq \sigma^2\,(1 + \sqrt{y})$ el estimador de λ no se separa del "montón".
    Para alguna transformación ortogonal $\mathbf{U}$

$$\mathbf{U\Sigma U}^T = \text{diag}(\lambda_1, \lambda_2, \ldots, \ \ldots \lambda_d, \sigma^2, \ldots, \ \ldots, \sigma^2)$$

  - Esos λ's son, en consecuencia, difíciles (o imposible) de detectar.

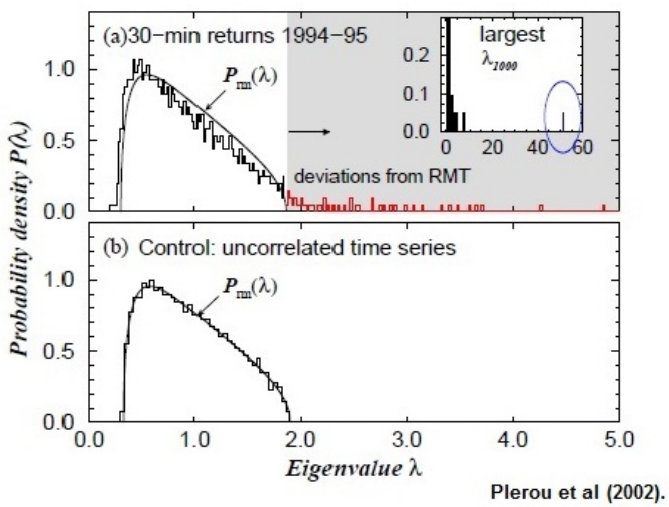
¿Cómo estimar la dimensión que importa?

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Uno puede (Plerou et al, Bouchaud etc.):

- Sacar el autovalor más grande y ver si el resto de los autovalores se ajustan a la distribución MP (gráfico inferior).
- Si esto sucede, la dimensión es igual a 1.
- En otro caso, sacamos el segundo autovalor más grande y repetimos.

Observemos nuevamente este gráfico:



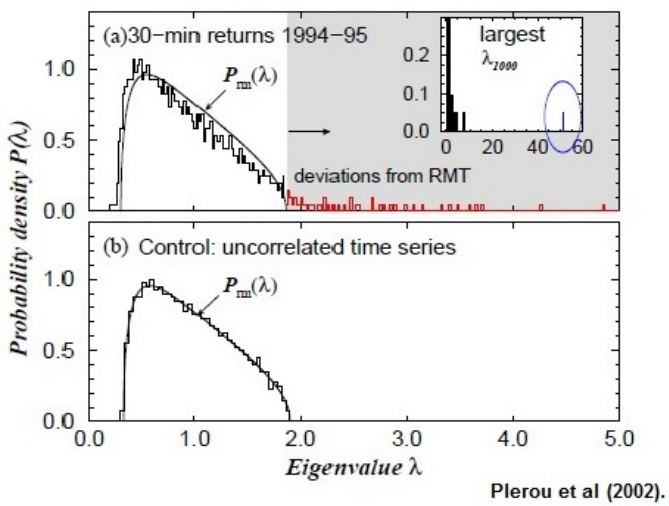
Plerou et al (2002).

¿Cómo estimar la dimensión que importa?

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Otro enfoque (Passemier, D. y Yao, J.F. (2012)):

- Separación entre los autovalores es menor en la parte correspondiente al ruido.
- La parte más estrecha corresponde a ruido, la más dispersa es la información.
- Chequear cuando el espaciado pasa de "ancho" a "estrecho".



Plerou et al (2002).

Objetivo de nuestro trabajo

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Hallar un test de cociente de máxima verosimilitud para determinar cuántas componentes importan.

Test de cociente de verosimilitud

Trabajos previos:

Introducción
Motivación

Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión

Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado

Conclusión

$$H_0 : d = 0 \quad \text{vs} \quad H_a : d > 0$$

Para $p < n$, el LRT es:

$$\text{LRT} = \frac{\hat{\lambda}_1 \dots \hat{\lambda}_p}{\left(\frac{1}{p}\sum_{i=1}^{p} \hat{\lambda}_i\right)^p}.$$

Test de cociente de verosimilitud

Trabajos previos:

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

$$H_0 : d = 0 \quad \text{vs} \quad H_a : d > 0$$

Para $p < n$, el LRT es:

$$\text{LRT} = \frac{\hat{\lambda}_1 \dots \hat{\lambda}_p}{\left(\frac{1}{p}\sum_{i=1}^{p} \hat{\lambda}_i\right)^p}.$$

- Un resultado asintótico clásico (Muirhead,1982) para p fijo:

$$-(n-1)\rho \log LRT \xrightarrow{d} \chi^2_f \quad \text{cuando } n \to \infty$$

El factor $\rho = \rho_n \to 1$ es un término de corrección para mejorar la tasa de convergencia.

Test de cociente de verosimilitud

Trabajos previos:

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

$$H_0 : d = 0 \quad \text{vs} \quad H_a : d > 0$$

Para $p < n$, el LRT es:

$$\text{LRT} = \frac{\hat{\lambda}_1 \dots \hat{\lambda}_p}{\left(\frac{1}{p}\sum_{i=1}^{p} \hat{\lambda}_i\right)^p}.$$

- Para el caso $p \to \infty$ con $p/n \to y \leq 1$ Jiang and Yang (2013) prueban que

$$\frac{\log(LRT) - \mu^*}{\sigma^*} \longrightarrow N(0,1)$$

y a partir de la distribución obtienen un test asintótico.

Test de cociente de verosimilitud

Trabajos previos:

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

$$H_d : \text{dimensión} = d \quad \text{vs} \quad H_a : \text{dimensión} > d$$

Para $p < n$, el LRT_d es:

$$LRT_d = \frac{\hat{\lambda}_{d+1} \dots \hat{\lambda}_p}{\left(\frac{1}{p-d} \sum_{i=d+1}^{p} \hat{\lambda}_i\right)^{p-d}}.$$

Test de cociente de verosimilitud

Trabajos previos:

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

$$H_d : \text{ dimensión } = d \quad \text{vs} \quad H_a : \text{dimensión } > d$$

Para $p < n$, el LRT_d es:

$$\text{LRT}_d = \frac{\hat{\lambda}_{d+1} \dots \hat{\lambda}_p}{\left(\frac{1}{p-d} \sum_{i=d+1}^{p} \hat{\lambda}_i\right)^{p-d}}.$$

- De nuevo, si p es fijo y $n \to \infty$, bajo H_d

$$-\rho \log LRT_d \to \chi^2_{(p-d+2)(p-d-1)/2},$$

(Lawley (1956), James (1969))

Test de cociente de verosimilitud

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

$$H_d : \text{ dimensión } = d \quad \text{vs} \quad H_a : \text{dimensión } > d$$

Para $p < n$, el LRT_d es:

$$LRT_d = \frac{\hat{\lambda}_{d+1} \dots \hat{\lambda}_p}{\left(\frac{1}{p-d} \sum_{i=d+1}^{p} \hat{\lambda}_i\right)^{p-d}}.$$

- En nuestro trabajo consideramos el caso $p \to \infty$ con $p/n \to y < 1$.
- Probamos que

$$\frac{\log(LRT_d) - \mu_{p,n,d}}{\sigma_{p,n,d}} \longrightarrow N(0,1)$$

- Test secuencial para estimar d.

Test de cociente de verosimilitud

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

¿Que hacemos si tenemos más variables que observaciones?

Test de cociente de verosimilitud

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

¿Que hacemos si tenemos más variables que observaciones?

LRT no existe, ni siquiera para el caso $\sigma^2 \mathbf{I}_p$ (hay muchos autovalores nulos).

Test de cociente de verosimilitud

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

¿Que hacemos si tenemos más variables que observaciones?

LRT no existe, ni siquiera para el caso $\sigma^2 \mathbf{I}_p$ (hay muchos autovalores nulos).

Sin embargo podemos:

Test de cociente de verosimilitud

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

¿Que hacemos si tenemos más variables que observaciones?

LRT no existe, ni siquiera para el caso $\sigma^2 \mathbf{I}_p$ (hay muchos autovalores nulos).

Sin embargo podemos:

- Invertir los roles de p y n.
- Definimos

$$\mathrm{LRT}_0^* = \frac{\hat{\lambda}_1 \dots \hat{\lambda}_n}{\left(\frac{1}{n}\sum_{i=1}^n \hat{\lambda}_i\right)^n}.$$

¿Qué significa invertir los roles de p y n?

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Sea $\widehat{\mathbf{\Sigma}}$ la covarianza muestral. Entonces, bajo H_0, $\mathbf{W} = m\widehat{\mathbf{\Sigma}} = \mathbf{Y}^T\mathbf{Y} \sim W_p(m, \mathbf{\Sigma})$ con $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$, donde $\mathbf{Y} \in \mathbb{R}^{m\times p} \sim N(0, \mathbf{I}_m \otimes \mathbf{\Sigma})$ (Muirhead, 1982).

¿Qué significa invertir los roles de p y n?

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Sea $\widehat{\mathbf{\Sigma}}$ la covarianza muestral. Entonces, bajo H_0, $\mathbf{W} = m\widehat{\mathbf{\Sigma}} = \mathbf{Y}^T\mathbf{Y} \sim W_p(m, \mathbf{\Sigma})$ con $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$, donde $\mathbf{Y} \in \mathbb{R}^{m\times p} \sim N(0, \mathbf{I}_m \otimes \mathbf{\Sigma})$ (Muirhead, 1982).
- Definimos $\widetilde{\mathbf{W}} = \mathbf{Y}\mathbf{Y}^T \sim W_m(p, \sigma^2\mathbf{I}_m)$ con $p > m$.

¿Qué significa invertir los roles de p y n?

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Sea $\widehat{\mathbf{\Sigma}}$ la covarianza muestral. Entonces, bajo H_0, $\mathbf{W} = m\widehat{\mathbf{\Sigma}} = \mathbf{Y}^T\mathbf{Y} \sim W_p(m, \mathbf{\Sigma})$ con $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$, donde $\mathbf{Y} \in \mathbb{R}^{m\times p} \sim N(0, \mathbf{I}_m \otimes \mathbf{\Sigma})$ (Muirhead, 1982).
- Definimos $\widetilde{\mathbf{W}} = \mathbf{Y}\mathbf{Y}^T \sim W_m(p, \sigma^2\mathbf{I}_m)$ con $p > m$.
- Los autovalores no nulos de $\mathbf{W}$ coinciden con los de $\widetilde{\mathbf{W}}$.

¿Qué significa invertir los roles de p y n?

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Sea $\widehat{\mathbf{\Sigma}}$ la covarianza muestral. Entonces, bajo H_0, $\mathbf{W} = m\widehat{\mathbf{\Sigma}} = \mathbf{Y}^T\mathbf{Y} \sim W_p(m, \mathbf{\Sigma})$ con $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$, donde $\mathbf{Y} \in \mathbb{R}^{m\times p} \sim N(0, \mathbf{I}_m \otimes \mathbf{\Sigma})$ (Muirhead, 1982).
- Definimos $\widetilde{\mathbf{W}} = \mathbf{Y}\mathbf{Y}^T \sim W_m(p, \sigma^2\mathbf{I}_m)$ con $p > m$.
- Los autovalores no nulos de $\mathbf{W}$ coinciden con los de $\widetilde{\mathbf{W}}$.
- Por tanto, LRT_0^* es el cociente de verosimilitud para $\widetilde{\mathbf{W}}$

¿Qué significa invertir los roles de p y n?

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Sea $\widehat{\mathbf{\Sigma}}$ la covarianza muestral. Entonces, bajo H_0, $\mathbf{W} = m\widehat{\mathbf{\Sigma}} = \mathbf{Y}^T\mathbf{Y} \sim W_p(m, \mathbf{\Sigma})$ con $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$, donde $\mathbf{Y} \in \mathbb{R}^{m\times p} \sim N(0, \mathbf{I}_m \otimes \mathbf{\Sigma})$ (Muirhead, 1982).
- Definimos $\widetilde{\mathbf{W}} = \mathbf{Y}\mathbf{Y}^T \sim W_m(p, \sigma^2\mathbf{I}_m)$ con $p > m$.
- Los autovalores no nulos de $\mathbf{W}$ coinciden con los de $\widetilde{\mathbf{W}}$.
- Por tanto, LRT_0^* es el cociente de verosimilitud para $\widetilde{\mathbf{W}}$
- La distribución de LRT_0^* se sigue de Jiang and Yang (2013), inviertiendo los roles de p y m.

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Esto nos motiva a definir LRT_d^* para esfericidad parcial, en el caso $p > n$ como:

$$\mathrm{LRT}_d^* = \frac{\hat{\lambda}_{d+1} \dots \hat{\lambda}_n}{\left(\frac{1}{n-d} \sum_{i=d+1}^{n} \hat{\lambda}_i\right)^{n-d}}.$$

- Para obtener un test de nivel asintótico α, necesitamos su distribución bajo H_d.
- Pero $\widetilde{\mathbf{W}}$ ya no es más una Wishart, sin embargo,

$$\widetilde{\mathbf{W}} = \sum_{i=1}^{d} \lambda_i \mathbf{Z}_{(i)} \mathbf{Z}_{(i)}^T + \sigma^2 \widetilde{\mathbf{Z}} \widetilde{\mathbf{Z}}^T$$

donde $\widetilde{\mathbf{Z}} \widetilde{\mathbf{Z}}^T \sim W_m(p-d, \mathbf{I}_m)$ independiente de $\mathbf{Z}_d = \left(\mathbf{Z}_{(1)}, \mathbf{Z}_{(2)}, \dots, \mathbf{Z}_{(d)}\right) \sim N(0, \mathbf{I}_m \otimes \mathbf{I}_d)$

Idea de la prueba

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Para el caso $p < n$:

Idea de la prueba

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Para el caso $p < n$:

1. Se estudian los momentos de LRT_d^t, para t cerca de 0.

Idea de la prueba

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Para el caso $p < n$:

1. Se estudian los momentos de LRT_d^t, para t cerca de 0.
2. Se prueba que $\log \mathbb{E}\left(\mathrm{LRT}_d^t\right) \underset{t \sim 0}{\sim} a_{p,n} t + b_{p,n}^2 \frac{t^2}{2} + o(1)$

Idea de la prueba

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Para el caso $p < n$:

1. Se estudian los momentos de LRT_d^t, para t cerca de 0.
2. Se prueba que $\log \mathbb{E}\left(\mathrm{LRT}_d^t\right) \underset{t \sim 0}{\sim} a_{p,n} t + b_{p,n}^2 \frac{t^2}{2} + o(1)$
3. Para p fijo, $a_{p,n}$ y $b_{p,n}$ fácil.

Idea de la prueba

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Para el caso $p < n$:

1. Se estudian los momentos de LRT_d^t, para t cerca de 0.
2. Se prueba que $\log \mathbb{E}\left(\mathrm{LRT}_d^t\right) \underset{t \sim 0}{\sim} a_{p,n} t + b_{p,n}^2 \frac{t^2}{2} + o(1)$
3. Para p fijo, $a_{p,n}$ y $b_{p,n}$ fácil.
4. Para $p, n \to \infty$ necesitamos límites de la función $\Gamma_p(\cdot)$.

Idea de la prueba

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Para el caso $p < n$:

1. Se estudian los momentos de LRT_d^t, para t cerca de 0.
2. Se prueba que $\log \mathbb{E}\left(\mathrm{LRT}_d^t\right) \underset{t \sim 0}{\sim} a_{p,n} t + b_{p,n}^2 \frac{t^2}{2} + o(1)$
3. Para p fijo, $a_{p,n}$ y $b_{p,n}$ fácil.
4. Para $p, n \to \infty$ necesitamos límites de la función $\Gamma_p(\cdot)$.
5. Luego $\frac{\log \mathrm{LRT}_d - a_{n,p}}{b_{n,p}} \to N(0,1)$

Idea de la prueba

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Para el caso $p < n$:

1. Se estudian los momentos de LRT_d^t, para t cerca de 0.
2. Se prueba que $\log \mathbb{E}\left(\mathrm{LRT}_d^t\right) \underset{t \sim 0}{\sim} a_{p,n}t + b_{p,n}^2 \frac{t^2}{2} + o(1)$
3. Para p fijo, $a_{p,n}$ y $b_{p,n}$ fácil.
4. Para $p, n \to \infty$ necesitamos límites de la función $\Gamma_p(\cdot)$.
5. Luego $\frac{\log \mathrm{LRT}_d - a_{n,p}}{b_{n,p}} \to N(0,1)$
6. En ambos caso, se usa fuertemente la densidad Wishart (cálculo de esperanza).

Idea de la prueba

Introducción

Motivación

Modelos Spiked

Johnstone (2001)

Estadística en alta dimensión

Estimando la dimensión

Enfoques previos

Nuestro objetivo: test LRT

"Information theoretic": estadístico penalizado

Conclusión

El caso $p > n$ (y $d > 0$)

Idea de la prueba

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

El caso $p > n$ (y $d > 0$)

1. $\widetilde{\mathbf{W}}$ ya no es Whishart.

Idea de la prueba

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

El caso $p > n$ (y $d > 0$)

1. $\widetilde{\mathbf{W}}$ ya no es Whishart.
2. Sin embargo $\mathrm{LRT}_d^* = \dfrac{|\widetilde{\boldsymbol{\Sigma}}|}{\left(\frac{1}{m-d}\sum_{i=d+1}^{m}\hat{\lambda}_i\right)^{m-d}} \dfrac{1}{\hat{\lambda}_1 \cdot \ldots \cdot \hat{\lambda}_d}$

Idea de la prueba

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

El caso $p > n$ (y $d > 0$)

1. $\widetilde{\mathbf{W}}$ ya no es Whishart.
2. Sin embargo $\mathrm{LRT}_d^* = \dfrac{|\widetilde{\mathbf{\Sigma}}|}{\left(\frac{1}{m-d}\sum_{i=d+1}^{m}\hat{\lambda}_i\right)^{m-d}} \dfrac{1}{\hat{\lambda}_1 \cdot \ldots \cdot \hat{\lambda}_d}$
3. $|\widetilde{\mathbf{\Sigma}}| = |\dfrac{\sigma^2}{m}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T||\mathbf{I}_d + \dfrac{1}{\sigma^2}\mathbf{\Lambda}_d^{1/2}\mathbf{Z}_d^T\left(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T\right)^{-1}\mathbf{Z}_d\mathbf{\Lambda}_d^{1/2}|$

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

1. A partir de la distibución asintótica obtenemos un test de nivel (asintótico) igual a α considerando la región de rechazo $\{LRT_d < c_\alpha\}$ donde $c_\alpha = \exp(C_\alpha)$ con $C_\alpha = \mu_{m,p,d} + z_\alpha \sigma_{m,p,d}$ y z_α es el cuantil α de la normal estándar.

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

1. A partir de la distibución asintótica obtenemos un test de nivel (asintótico) igual a α considerando la región de rechazo $\{LRT_d < c_\alpha\}$ donde $c_\alpha = \exp(C_\alpha)$ con $C_\alpha = \mu_{m,p,d} + z_\alpha \sigma_{m,p,d}$ y z_α es el cuantil α de la normal estándar.
2. Pero C_α depende de los verdaderos parámetros y, por lo tanto, debemos reemplazarlos por funciones de los valores muestrales.

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

1. A partir de la distibución asintótica obtenemos un test de nivel (asintótico) igual a α considerando la región de rechazo $\{LRT_d < c_\alpha\}$ donde $c_\alpha = \exp(C_\alpha)$ con $C_\alpha = \mu_{m,p,d} + z_\alpha \sigma_{m,p,d}$ y z_α es el cuantil α de la normal estándar.
2. Pero C_α depende de los verdaderos parámetros y, por lo tanto, debemos reemplazarlos por funciones de los valores muestrales.
3. σ^2 se puede reemplazar por su estimador consistente $\hat{\sigma}^2 = \sum_{i=d+1}^{p} \hat{\lambda}_i/(p-d)$.

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

1. A partir de la distibución asintótica obtenemos un test de nivel (asintótico) igual a α considerando la región de rechazo $\{LRT_d < c_\alpha\}$ donde $c_\alpha = \exp(C_\alpha)$ con $C_\alpha = \mu_{m,p,d} + z_\alpha \sigma_{m,p,d}$ y z_α es el cuantil α de la normal estándar.
2. Pero C_α depende de los verdaderos parámetros y, por lo tanto, debemos reemplazarlos por funciones de los valores muestrales.
3. σ^2 se puede reemplazar por su estimador consistente $\hat{\sigma}^2 = \sum_{i=d+1}^{p} \hat{\lambda}_i/(p-d)$.
4. Con los λ_i, debemos considerar la transición de fase y su sesgo.

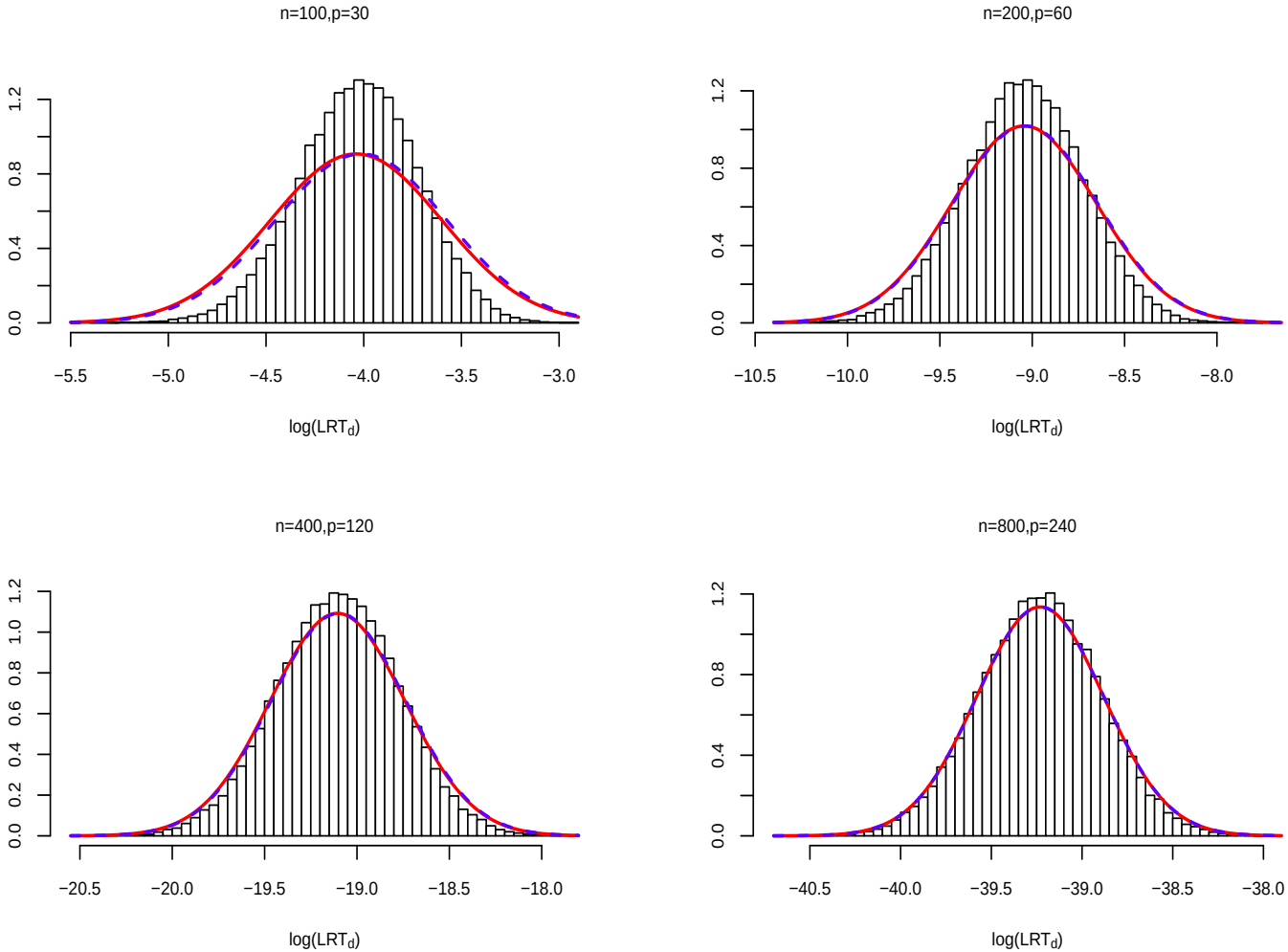
Simulación: Distribución asintótica

Introducción
Motivación

Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión

Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado

Conclusión



$$d = 4,\ \lambda = [7, 6, 5, 4],\ \sigma^2 = 1,\ p/n = 0.3.$$

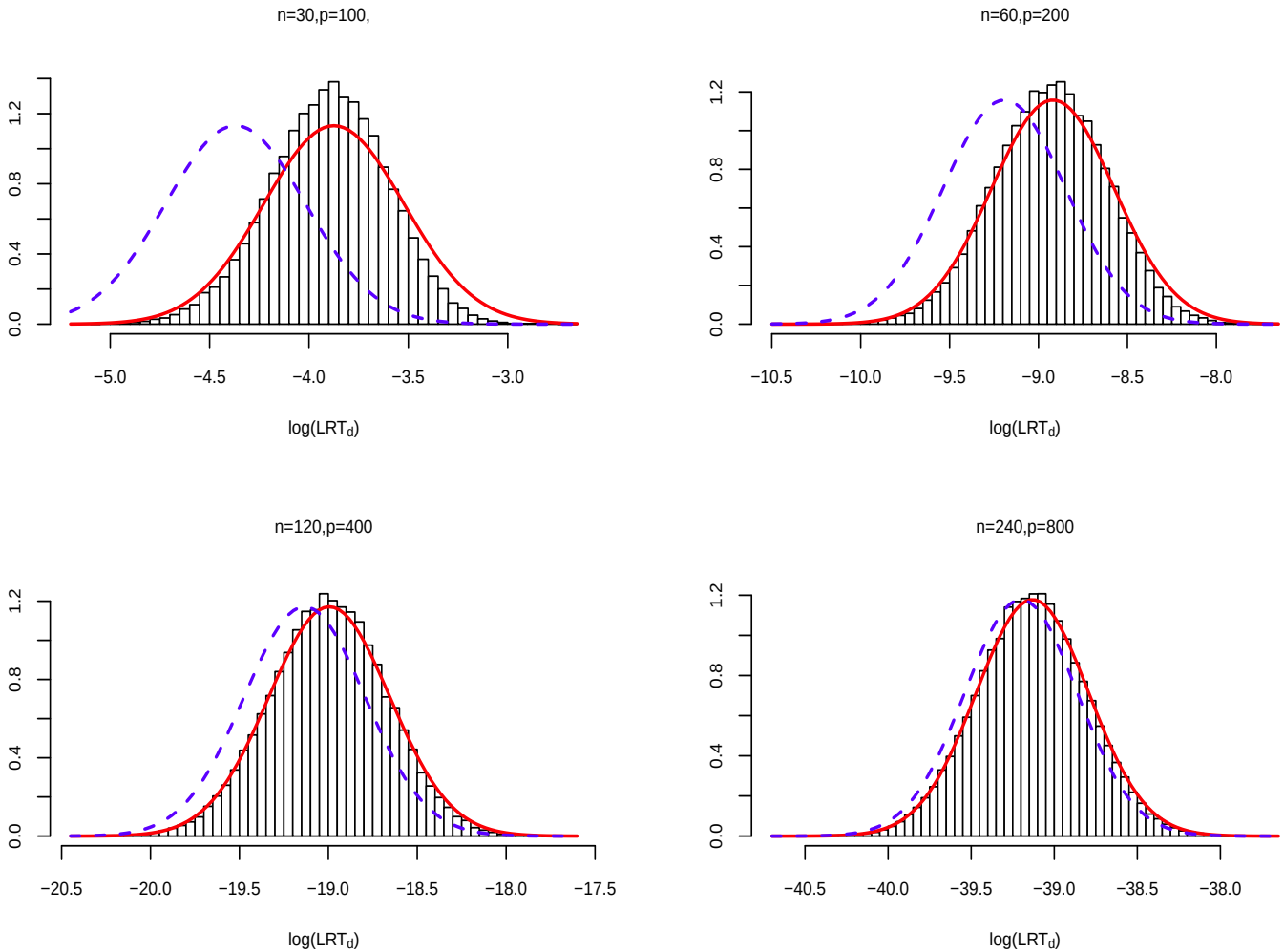
Simulación: Distribución asintótica

Introducción
Motivación

Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión

Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado

Conclusión



$$d = 4,\ \lambda = [7, 6, 5, 4],\ \sigma^2 = 1,\ p/n = 1/0.3.$$

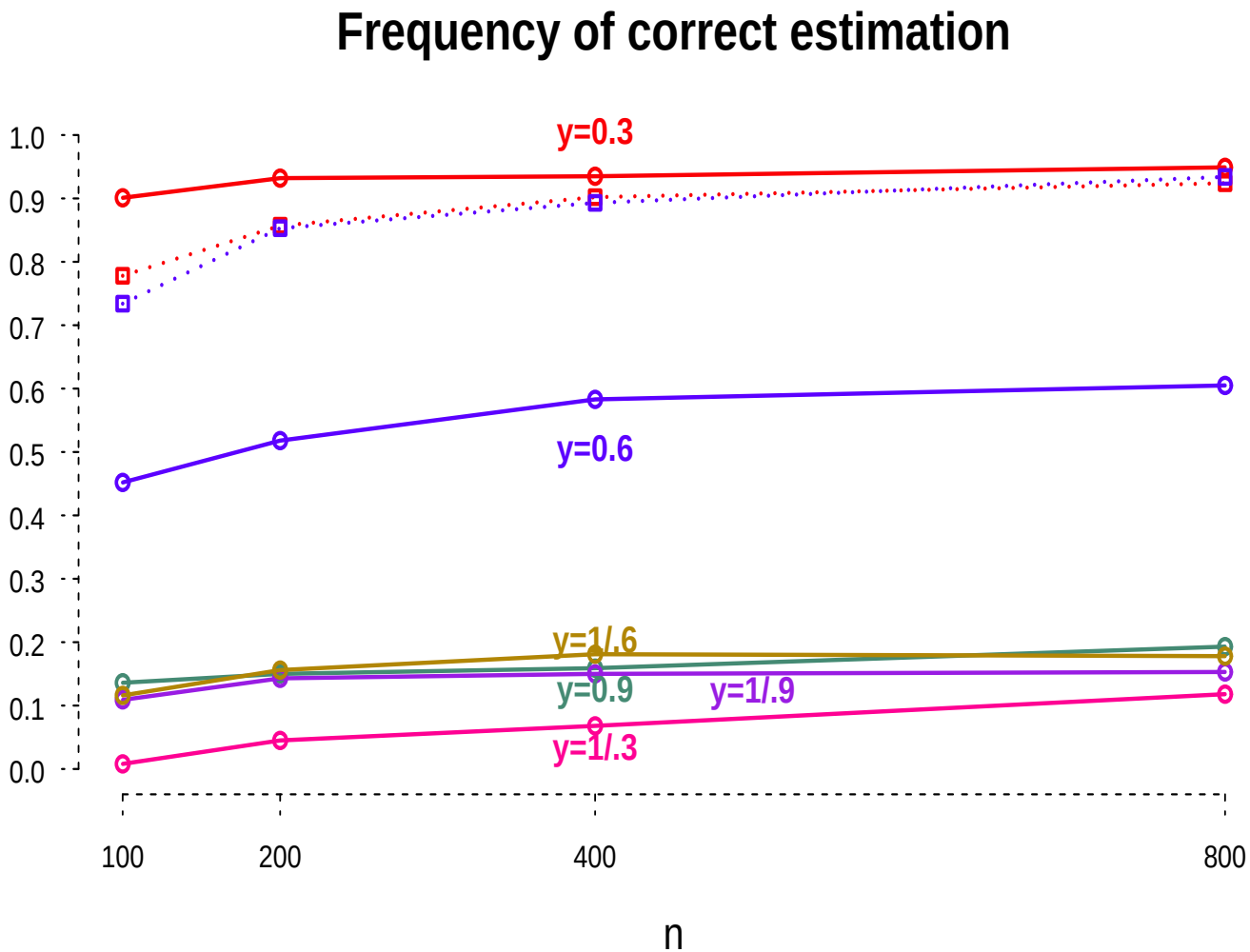
Simulación: Estimación de la dimensión

Introducción
Motivación

Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión

Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado

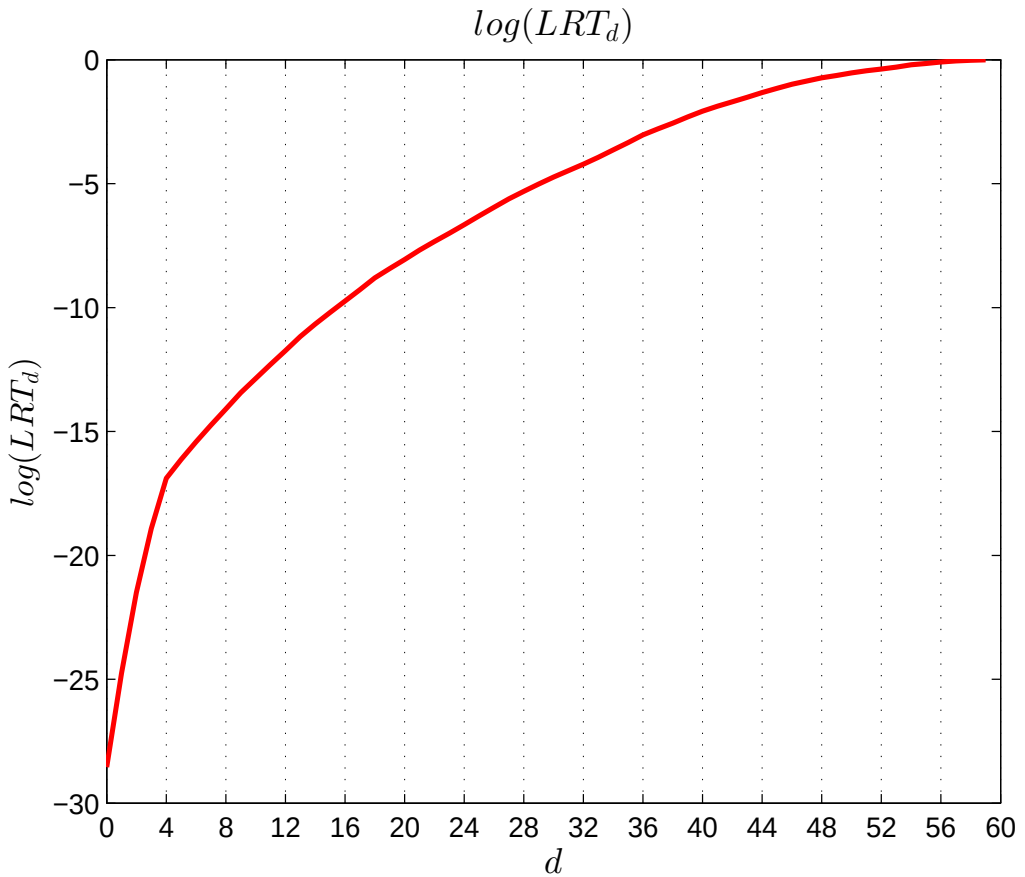
Conclusión



Hacia la versión penalizada

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Consideremos nuevamente $d = 4$, $\lambda = [7, 6, 5, 4]$, $\sigma^2 = 1$
Graficamos la función $\log(LRT_d)$:



Hacia la versión penalizada

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

Estudio de los incrementos:

- Después de $d = 4$, (en el límite cuando $p, n \to \infty$) sólo dependen de p/n.
- Antes de $d = 4$, dependen de las estimaciones $\hat{\lambda}_i$.
- Es difícil detectar el quiebre.
- La penalización tiene en cuenta este cambio.
- Idea para penalizar: en cada paso restar el menor incremento.

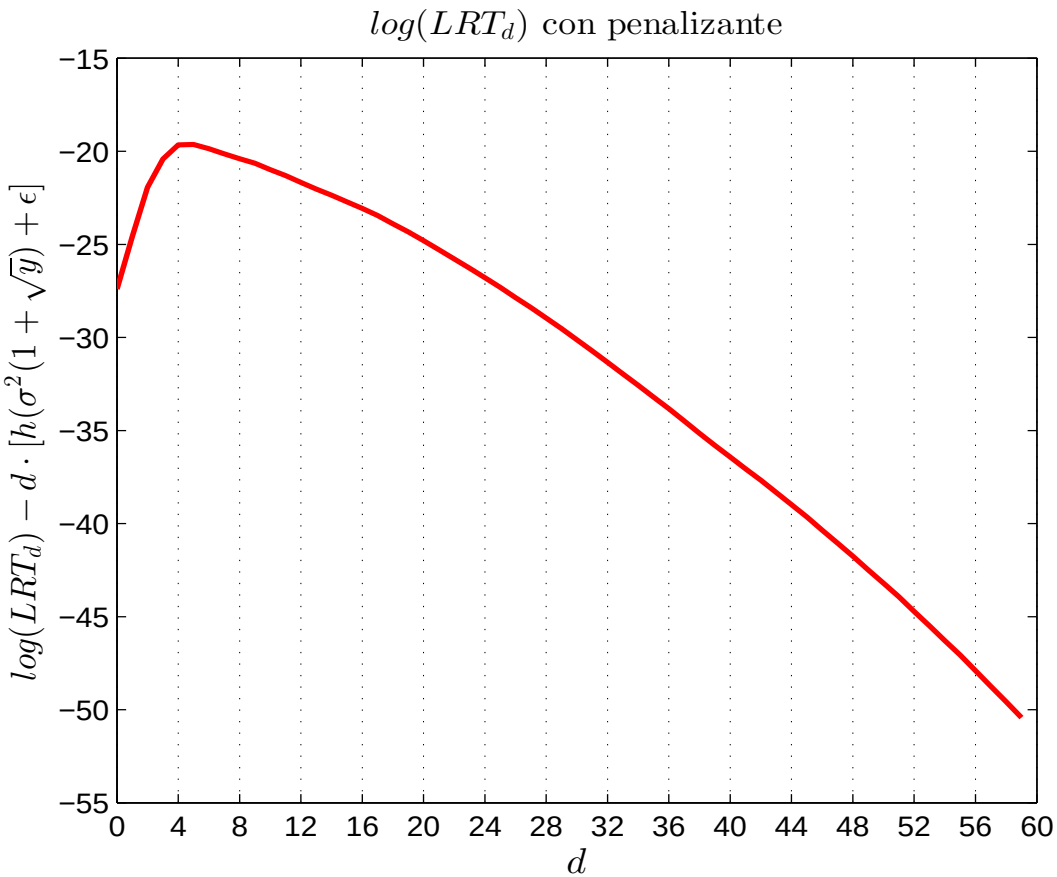
Hacia la versión penalizada

Introducción
Motivación

Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión

Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado

Conclusión



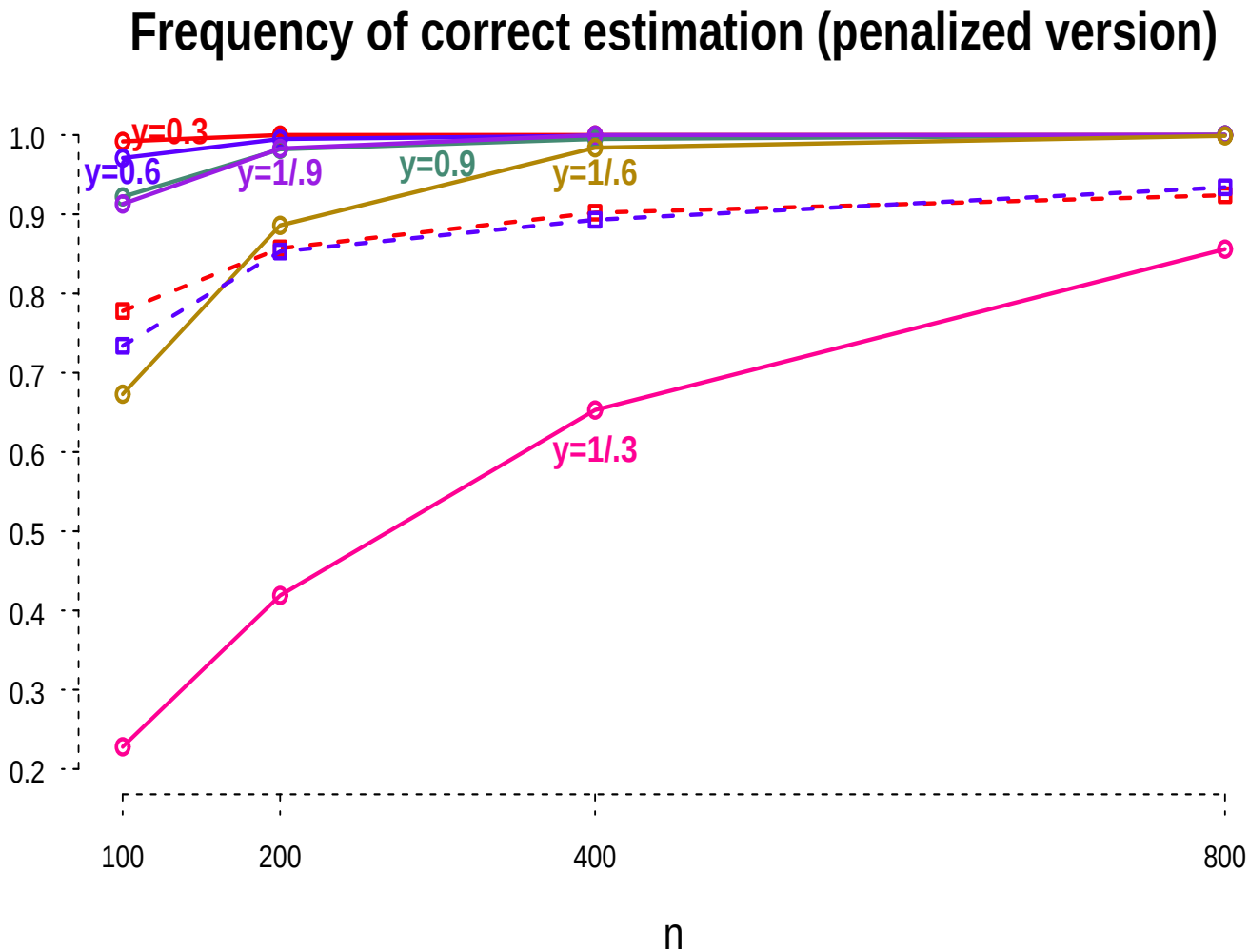
Simulación: Estimación de la dimensión

Introducción
Motivación

Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión

Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado

Conclusión



Conclusión

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Para el caso $p < n$ y $p, n \to \infty$, estudiamos la distribución asintótica del LRT_d para esfericidad parcial para el caso de modelos de covarianza spiked introducidos por Johnstone (2001).

Conclusión

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Para el caso $p < n$ y $p, n \to \infty$, estudiamos la distribución asintótica del LRT_d para esfericidad parcial para el caso de modelos de covarianza spiked introducidos por Johnstone (2001).
- Para $p > n$ y $p, n \to \infty$ obtenemos también la distribución asintótica del estadístico LRT_d invirtiendo los roles de p y n.

Conclusión

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Para el caso $p < n$ y $p, n \to \infty$, estudiamos la distribución asintótica del LRT_d para esfericidad parcial para el caso de modelos de covarianza spiked introducidos por Johnstone (2001).
- Para $p > n$ y $p, n \to \infty$ obtenemos también la distribución asintótica del estadístico LRT_d invirtiendo los roles de p y n.
- Conocer estas distribuciones nos permite desarrollar un test para elegir la dimensión de la parte spike .

Conclusión

Introducción
Motivación
Modelos Spiked
Johnstone (2001)
Estadística en alta dimensión
Estimando la dimensión
Enfoques previos
Nuestro objetivo: test LRT
"Information theoretic": estadístico penalizado
Conclusión

- Para el caso $p < n$ y $p, n \to \infty$, estudiamos la distribución asintótica del LRT_d para esfericidad parcial para el caso de modelos de covarianza spiked introducidos por Johnstone (2001).
- Para $p > n$ y $p, n \to \infty$ obtenemos también la distribución asintótica del estadístico LRT_d invirtiendo los roles de p y n.
- Conocer estas distribuciones nos permite desarrollar un test para elegir la dimensión de la parte spike .
- Mediante el estudio del comportamiento de la distribución del estadístico LRT_d para valores por debajo y por encima de la verdadera dimensión del spike, podemos modificar el estimador de la dimensión y probar la consistencia del mismo.

Muchas gracias!!