MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula DAD

9 de março de 2026

Sumário

1. Teoria Clássica versus Alta Dimensão
2. Exemplos em que resultados clássicos falham
3. Considerações sobre os exemplos

Preliminares

- A partir de 1950 surge o interesse por dados de alta dimensão (DAD). Trabalhos de Rao, Wigner, Kolmogorov, Huber etc.
- Nas últimas duas décadas o interesse em DAD aparece com a análise de dados extremamente grandes, com dim$=O(n)$ ou maior, em que $n =$ número de amostras.
- Em tais casos, a teoria assintótica clássica pode não funcionar.
- O desenvolvimnento em DAD tem relações com MV, otimização, análise numérica, análise funcional, teoria da informação, teoria da aproximação e teoria de probabildades.

Preliminares

- A teoria clássica em probabilidade e estatística fornecem afirmações aplicáveis a uma classe fixa de modelos, parametrizada por um índice n (o tamanho amostral) que pode crescer.
- O exemplo canônico de tal teoria é a LGN, que garante que a média amostral $\hat{\mu}_n$ de uma v.a X com média μ e variância finita σ^2, converge em probabildade para μ, ou seja, $\hat{\mu}_n$ é um estimador consistente de μ.
- Uma afirmação mais refinada é dada pelo TLC, que garante que o desvio re-escalado $\sqrt{n}(\hat{\mu}_n - \mu)$ converge, em distribuição, para uma v.a gaussiana, com matriz de covariância $\Sigma = \text{Cov}(X_i)$.
- Essas duas afirmações teóricas são subjacentes a uma grande classe de estimadores, em parte garantindo sua normalidade assintótica.

Preliminares

- No enfoque clássico, a dimensão d do espaço dos dados é suposta fixa.
- Motivação para DAD: Suponha que tenhamos $n = 1000$ amostras de um modelo em $d = 500$ dimensões. A teoria que requer que $n \to \infty$ com d fixa, fornecerá predições úteis?
- Essa questão não pode ser respondida definitivamente sem detalhes adicionais sobre o modelo sob consideração.
- Alguns fatos essenciais: a) dados aparecendo atualmente têm $d = O(n)$ ou $d > O(n)$; b) para muitas aplicações, a teoria que supõe "n grande e d fixa" falha em fornecer predições úteis; c) métodos clássicos podem falhar dramaticamente em regimes DAD.

Exemplo 1: Análise discriminante linear (LDA)

- Considere o problema do teste de hipótese binário:

$$\mathbf{x} \in \mathbb{R}^d \sim \mathbb{P}_1 \quad \text{versus } \mathbf{x} \in \mathbb{R}^d \sim \mathbb{P}_2.$$

- Quandp $\mathbb{P}_1$ e $\mathbb{P}_2$ são conhecidos, uma regra de decisão natural é baseada em limitar (threshoding) a razão de versossimilhanças $\log \frac{\mathbb{P}_2(\mathbf{x})}{\mathbb{P}_1(\mathbf{x})}$. Variando esse limiar permite um balanço entre os dois tipos de erros: decidir por $\mathbb{P}_1$ quando $\mathbb{P}_2$ for verdadeiro e vice-versa.
- O Lema de Neyman-Pearson garante que essa família de regras de decisões, possivelmente com aleatorização, é ótima, no sentido que elas traçam a curva dando o melhor compromisso (trade-off) entre os dois tipos de erros.

Exemplo 1: LDA

- Suponha, por exemplo, que $\mathbb{P}_1 \sim N(\mu_1, \Sigma)$ e $\mathbb{P}_2 \sim N(\mu_2, \Sigma)$. A log-verossimilhança reduz-se à estatística linear

$$\begin{aligned} \Psi(\mathbf{x}) &= \left\langle \mu_1 - \mu_2, \Sigma^{-1}\left(\mathbf{x} - \frac{\mu_1 + \mu_2}{2}\right)\right\rangle \\ &= (\mu_1 - \mu_2)^\top \Sigma^{-1}\left(\mathbf{x} - \frac{\mu_1 + \mu_2}{2}\right). \end{aligned} \tag{1}$$

- A regra de decisão ótima é baseada em obter um limiar (threshold) para essa estatística. Se as duas classes são igualmente prováveis, a probabilidade de classificação incorreta é dada por

$$\text{Err}(\Psi) = \frac{1}{2}\mathbb{P}_1[\Psi(X^{'}) \leq 0] + \frac{1}{2}\mathbb{P}_2[\Psi(X^{''}) > 0],$$

com $X^{'} \sim \mathbb{P}_1$, $X^{''} \sim \mathbb{P}_2$.

Exemplo 1: LDA

- Dada a normalidade, obtemos

$$\text{Err}(\Psi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\gamma/2} e^{-t^2/2} dt, \tag{2}$$

em que $\gamma = \sqrt{(\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2)}$ e a integral é igual a $\Phi(-\gamma/2)$.

- Na prática, as distribuições das duas classes não são conhecidas, mas se obervarmos $(\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}) \sim \mathbb{P}_1$ e $(\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_1+n_2)}) \sim \mathbb{P}_2$, uma abordagem natural é usar as amostras para estimar as quantidades de interesse.
- No nosso caso,

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_i, \quad \hat{\mu}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} \mathbf{x}_i,$$

$$\hat{\Sigma} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\mathbf{x}_i - \hat{\mu}_1)(\mathbf{x}_i - \hat{\mu}_1)^\top + \frac{1}{n_2 - 1} \sum_{i=n_1+1}^{n_1+n_2} (\mathbf{x}_i - \hat{\mu}_2)(\mathbf{x}_i - \hat{\mu}_2)^\top$$

Exemplo 1: LDA

- Substituindo em (1) obtemos a função discriminante linear de Fisher

$$\hat{\Psi}(\mathbf{x}) = (\hat{\mu}_1 - \hat{\mu}_2)^\top \hat{\Sigma}^{-1} \left(\mathbf{x} - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} \right). \tag{3}$$

- Se $n_i > d$, então $\hat{\Sigma}^{-1}$ existe.
- $\mathrm{Err}(\hat{\Psi})$ é obtido substituindo $\hat{\Psi}$ em (2).
- Em 1960, Kolmogorov analizou uma versão mais simples de (3), com $\Sigma = I$:

$$\hat{\Psi}_{\mathrm{id}}(\mathbf{x}) = (\hat{\mu}_1 - \hat{\mu}_2)^\top \left(\mathbf{x} - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} \right). \tag{4}$$

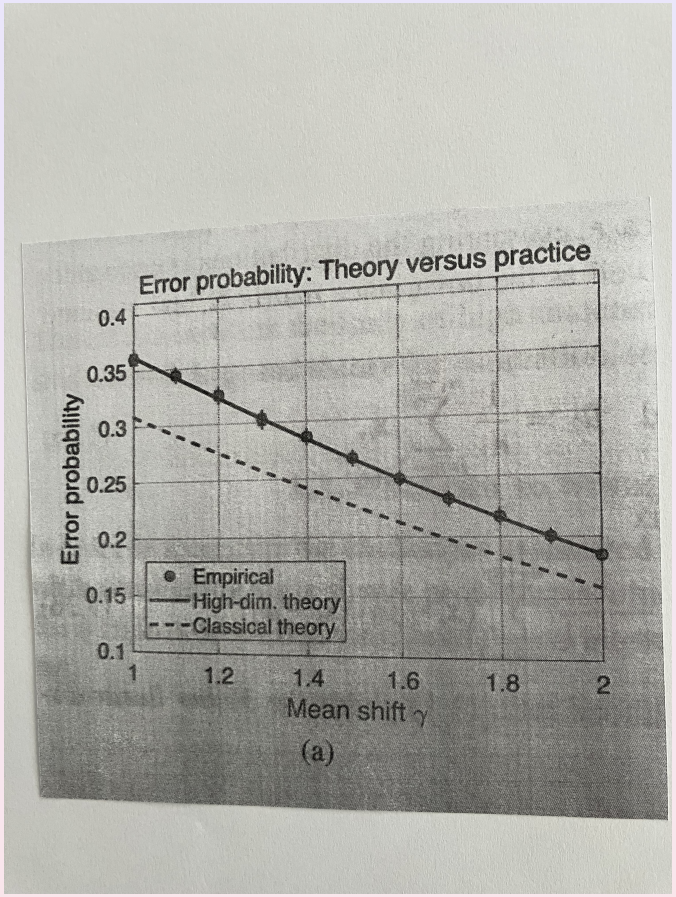
Exemplo 1: LDA

- Trabalhando sob a suposição de dados Gaussianos, ele analisou o comportamento deste método sob uma forma de limite na qual a tripla $(n_1, n_2, d) \to \infty$, com $d/n_i \to \alpha > 0$, $i = 1, 2$ e $||\mu_1 - \mu_2||_2 \to \gamma > 0$.
- Sob esse tipo de escalamento de AD, ele mostrou que

$$\text{Err}(\hat{\Psi}_{id}) \xrightarrow{P} \Phi\left(-\frac{\gamma^2}{2\sqrt{\gamma^2 + 2\alpha}}\right) \tag{5}$$

- Se $d/n_i \to 0$, obtemos $\Phi(-\gamma/2)$, como previsto por (2). Mas, se $d/n_i \to \alpha > 0$, o erro é estritamente maior do que o previsto pela previsão clássica, pois $\gamma^2/(2\sqrt{\gamma^2 + 2\alpha}) \to 0$.
- Veja a Figura 1. (a) Previsão HD, com $\alpha = 0,5$ mostra concordância excelente com o comportamnto na prática e a previsão clássica subestima o comportamento na prática; (b) Previsão clássica não depende de α, logo é constante e comportamento na prática coincide com previsão HD.

Exemplo 1: LDA



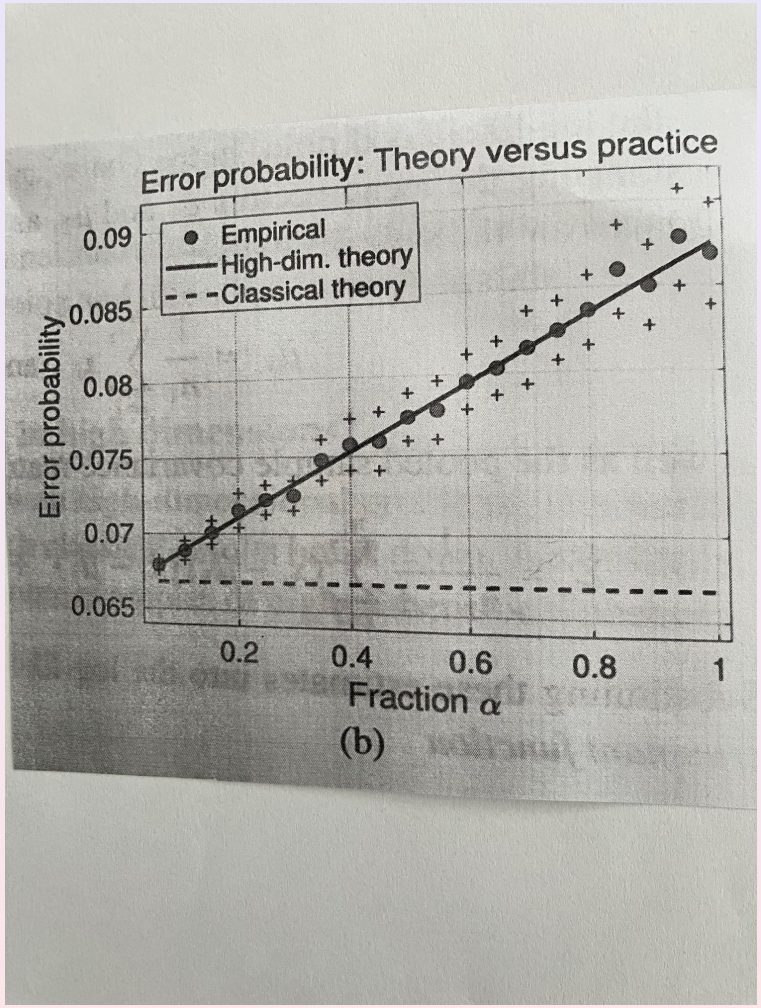


Figura 1: (a) $\gamma \in [1, 2], \alpha = 0.5, d = 400 (n_1 = n_2 = 800)$; (b) $\alpha \in [0, 1], d = 400, \gamma = 2$. Circles correspond to empirical performance averaged over 50 trials.

Exemplo 2: Estimação da matriz de covariâncias

- Como antes, temos $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \sim$ i.i.d. $\mathbb{P} \in \mathbb{R}^d$ e queremos estimar $\Sigma = \text{Cov}(\mathbf{X})$.
- Um estimador natural é a **matriz de covariâncias amostrais**

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top, \tag{6}$$

  e temos que $E(\hat{\Sigma}) = \Sigma$.
- Uma análise clássica considera o comportamento de $\hat{\Sigma}$ quando n cresce e d permance fixo. Para medir o comportamento da distância entre Σ e $\hat{\Sigma}$, usamos a norma ℓ_2,

$$||\hat{\Sigma} - \Sigma||_2 = \sup_{u \neq 0} \frac{||(\hat{\Sigma}) - \Sigma)u||_2}{||u||_2}. \tag{7}$$

Exemplo 2: Estimação da matriz de covariâncias

- Se $S_n = n\hat{\Sigma}$, então $S_n \sim \text{Wishart}_d(n, \Sigma)$, isto é, S_n tem distribuição de Wishart com n graus de liberdade e matriz escala Σ.
- Sob condições de momento não muito fortes, usando a LGN, pode-se mostrar que $||\hat{\Sigma} - \Sigma||_2 \to 0$ q.c., quando $n \to \infty$. Logo, $\hat{\Sigma}$ é um estimador fortemente consistente de Σ.
- Então, se $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ são os autovalores de Σ e $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_d$ são os autovalores de $\hat{\Sigma}$, então $\hat{\lambda}_j \to \lambda_j$ q.c., $j = 1, \ldots, d$.
- Além disso, $\sqrt{n}(\hat{\lambda}_k - \lambda_k) \to N(0, 2\lambda_k)$, para cada autovalor λ_k.
- Pode-se também, obter a distribuição normal assintótica

$$\sqrt{n}(\text{vec}(\hat{\Sigma}) - \text{vec}(\Sigma)) \to N_{d^2}(0, \Omega),$$

onde

$$\Omega = \text{Cov}(\text{vec}(\mathbf{X}\mathbf{X}^\top)).$$

Exemplo 2: Estimação da matriz de covariâncias

- A questão que surge é: o que acontece se permitirmos $d \to \infty$? Para averiguar isso, considere a sequência de problemas $(\hat{\Sigma}, \Sigma)$ indexada pelo par (n, d) e suponha que que permitimos que n e d crescem, mas com $d/n = \alpha \in (0, 1)$.
- A lei de Marchenko-Pastur (MP) nos diz que, sob condições moderadas, temos a convergência dos autovalores da matriz $\hat{\Sigma}$ para uma densidade estritamente positiva, com suporte em $[\alpha_-, \alpha_+]$, onde

$$\alpha_- = (1 - \sqrt{\alpha})^2, \quad \alpha_+ = (1 + \sqrt{\alpha})^2. \tag{8}$$

- Por exemplo, se $\alpha = 0,2$, $\alpha_- = 0,305$ e $\alpha_+ = 2,09$.
- A densidade assintótica, para $0 < \alpha < 1$, é dada por

$$f_\alpha(x) = \frac{1}{2\pi\alpha x}\sqrt{[(\alpha_+ - x)(x - \alpha_-)]_+}.$$

Exemplo 2: Estimação da matriz de covariâncias

- Vamos supor que $\Sigma = I_d$ e cada $\mathbf{X}_i \sim N(0, I_d)$, $i = 1, \ldots, n$. Usando essas n amostras geramos a matriz de covariâncias amostral (6) e calculamos seu vetor de autovalores $\gamma(\hat{\Sigma})$, arranjados em ordem não-decrescente

$$\gamma_{max}(\hat{\Sigma}) = \gamma_1(\hat{\Sigma}) \geq \gamma_2(\hat{\Sigma}) \geq \cdots \geq \gamma_d(\hat{\Sigma}) = \gamma_{min}(\hat{\Sigma}) \geq 0.$$

- Na Figura 2 mostramos os gráficos do vetor $\gamma(\hat{\Sigma}) \in \mathbb{R}^d$ de autovalores. O caso (a) corresponde a $(n, d) = (4000, 800)$ ou $\alpha = 0,2$. O caso (b) corresponde a $(n, d) = (4000, 2000)$ ou $\alpha = 0,5$.
- Em ambos os casos, obtemos $\gamma_{max}(\hat{\Sigma}) \approx (1 + \sqrt{\alpha})^2$, consistente com a teoria (veja $f_\alpha(\mathbf{x})$ acima).
- Se a matriz de covariância amostral estivesse convergindo para I_d, então o vetor de autovalores deveria convergir para um vetor de 1's e o histograma correspondente deveria se concentrar ao longo de 1, mas isso não acontece, tendo formas diferentes dependendo de $\alpha = d/n$.

Exemplo 2: Estimação da matriz de covariâncias

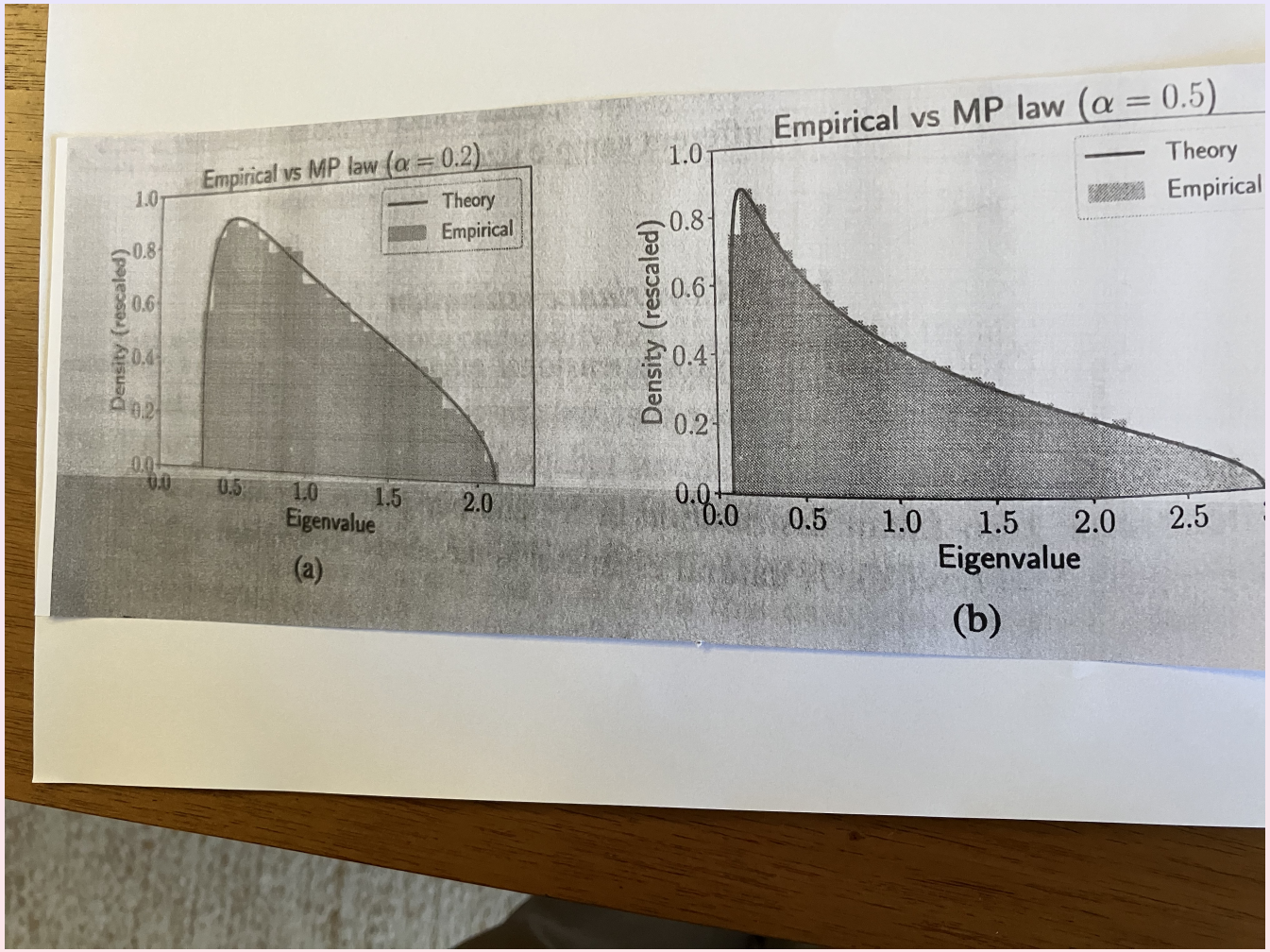


Figura 2: (a) $(n, d) = (4000, 800)$ ou $\alpha = 0,2$;
(b) $(n, d) = (4000, 2000)$ ou $\alpha = 0,5$.

Exemplo 2: Estimação da matriz de covariâncias

- A lei de M-P é uma afirmação assintótica, que permite n e p divergirem para $+\infty$.
- Se procurarmos por um resultado não assintótico, ou seja, que vale para todas as escolhas dos pares (n, d), podemos tentar obter limites inferior e superior. Pode-se provar que

$$P(\gamma_{max}(\hat{\Sigma}) \geq (1 + \sqrt{d/n} + \delta)^2 \leq e^{-n\delta^2/2}, \quad \text{para todo } \delta \geq 0. \qquad (9)$$

  com um resultado análogo para um limite inferior do $\gamma_{min}(\hat{\Sigma})$.
- O resultado (9) mostra que a probabilidade de que o autovalor máximo se desvie de $(1 + \sqrt{d/n})^2$ decai exponencialmente com n.
- Esse resultado também pode ser usado para mostrar que $\hat{\Sigma}$ é um estimador *operator-norm-consistent* de Σ, desde que $d/n \to 0$.
- Por exemplo, se $d/n = 0,2$, $n = 4000$, $\delta = 0,1$ a probabilidade em (9) é menor ou igual a $e^{-20} = 2.061154 \cdot 10^{-9}$.

V. A. Marchenko and L. A. Pastur (1967). Distribution of eigenvalues for some sets of random matrices, *Math. USSR Sbornik 1*, 457-483.

Exemplo 3: Regressão não paramétrica

- Os efeitos de AD em problemas de regressão podem ser ainda mais dramáticos.
- O modelo a ser considerado é

$$y = f(\mathbf{x}), \tag{10}$$

em que $\mathbf{x} \in [0,1]^d$, $f : [0,1]^d \to \mathbb{R}$ é uma função pertencente a algum espaço de funções.

- Queremos f de modo a minimizar $E[Y - f(\mathbf{X})]^2$, que resulta $f(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$.
- Na prática, a distribuição conjunta P_{XY} de $(\mathbf{X}, Y)$ não é conhecida, de modo que obter f diretamente não é possível.
- Temos uma amostra $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n,$, escolhida de modo i.i.d. de P_{XY} e nosso objetivo é encontrar $\hat{f}$ para a qual o EQM

$$||\hat{f} - f|||_{L_2}^2 = E_X[(\hat{f}(\mathbf{X}) - f(\mathbf{X}))^2] \tag{11}$$

seja o menor possível.

Exemplo 3: Regressão não paramétrica

- Mas esse problema torna-se extremamente difícil em AD, o chamado *curse of dimensionality*.
- A questão que se põe é: Quantas amostras devemos esolher como uma função de d?
- Por exemplo: suponha $\mathbf{X} \sim \mathcal{U}[0,1]^d$, de modo que P_X é a distribuição uniforme em $[0,1]^d$. Se pudermos gerar um bom estimador de f, baseado em $\mathbf{X}_1, \ldots, \mathbf{X}_n$, então um vetor típico $\mathbf{X}' \in [0,1]^d$ seria relativamente próximo de pelo menos uma de nossas amostras.
- Para formalizar esta noção, podemos considerar a quantidade

$$\rho_\infty(n,d) = E_{X',X}[\min_{1 \leq i \leq n} ||\mathbf{X}' - \mathbf{X}_i||_\infty], \tag{12}$$

que mede a distância média entre uma amostra independente $\mathbf{X}'$ extraída de $\mathcal{U}[0,1]^d$ e nosso conjunto de dados $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Exemplo 3: Regressão não paramétrica

- Posto de outra maneira: quantas amostras (n) precisamos coletar como uma função de d de modo a assegurar que $\rho_\infty(n, d) < \delta$, sendo δ um limiar?
- Por exemplo, se $\delta = 1/3$ nas simulações de (12) (20 réplicas), suponha que $d/n \to \alpha > 0$, digamos $\alpha = 0,5$, de modo que $n = 2d$.
- Na Figura 3 temos o gráfico do log(dist. min.) em função de d, em duas situações, $n = 2d$ e $n = d^2$.
- Notando que $log(1/3) \approx -1,1$, vemos que a distância mínima pode ser consideravelmente maior do que $1/3$.
- De fato, deveríamos tomar um tamanho amostral exponencialmente grande de maneira a assegurar que $\rho_\infty(n, d) < \delta$, à medida que d cresce.
- Pode-se provar que

$$\log \rho_\infty(n, d) \geq \log \frac{d}{2(d+1)} - \frac{\log n}{d}, \tag{13}$$

do que ressulta que deveríamos ter $n > (1/\delta)^d$ para se ter o limite superior acima.

Exemplo 3: Regressão não paramétrica

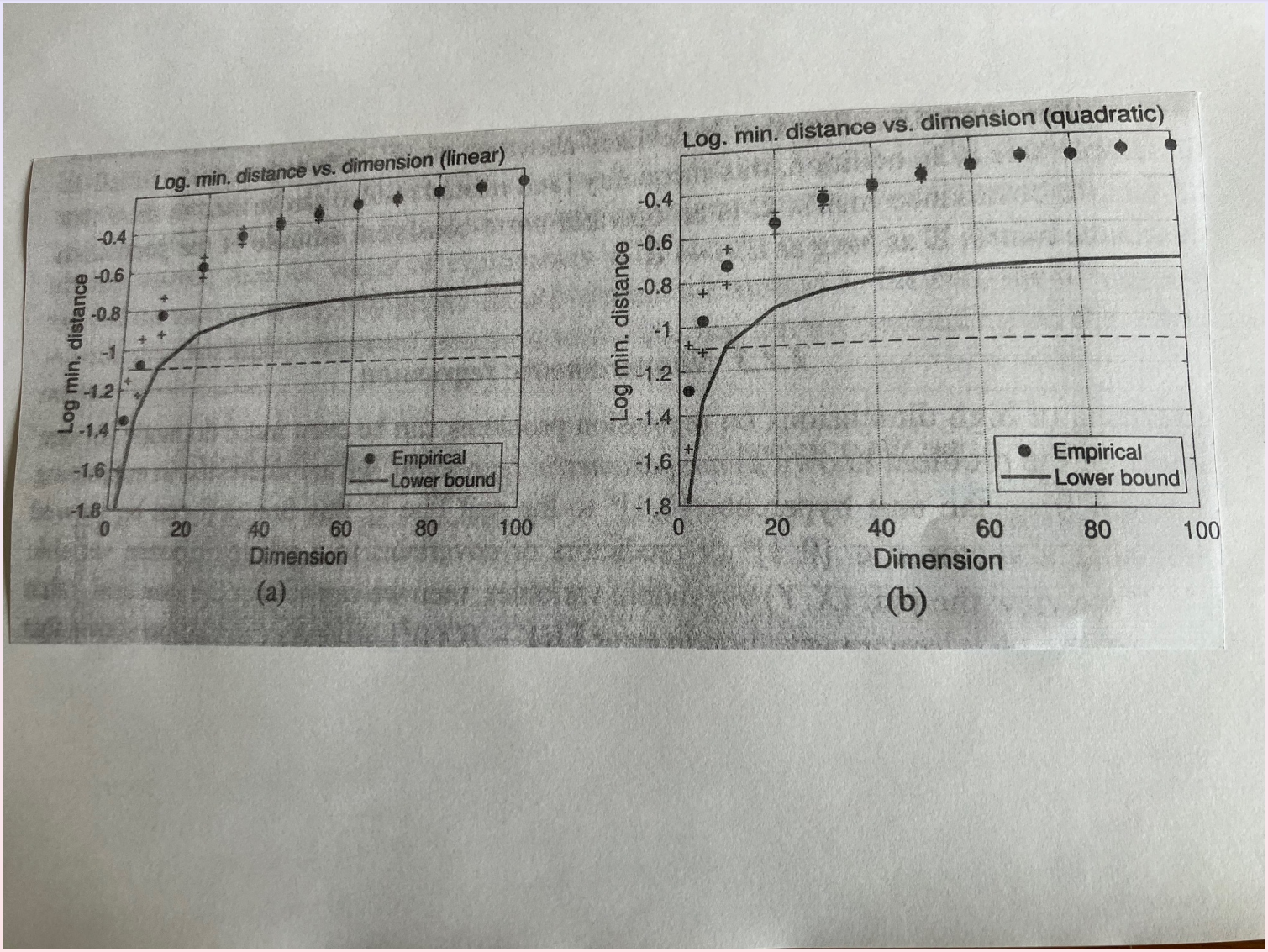


Figura 3: (a) $n = 2d$; (b) $n = d^2$.

Considerações Finais

- Os fenômenos de AD descritos são todos inevitáveis.
- Para o problema de classificação descrito, se a razão d/n permence estritamente acima de zero, então não é possível alcançar a taxa de classificação ótima.
- Para o problema de estimação da matriz de covariância, não há estimador consistente na norma ℓ_2 quando d/n permanece limitada acima de zero.
- Finalmente, para o problema de regressão não paramétrica, dado o objetivo de estimar uma regressão diferenciável para f, nenhum procedimento consistente é possível, a menos que o tamanho da amostra cresça exponencialmente com a dimensão d.
- Todas essas afirmações podem ser tornadas rigorosas com noções de entropia e estimação minimax (veja Wainwright (2019)).

Como ajudar quando temos DAD?

- Dado o exposto, o que pode ajudar quando temos DAD?
- Essencialmente, podemos apenas esperar que os dados possam ter alguma forma de estrutura de dimensão menor, que torne mais simples a solução do problema.
- A maioria das técnicas estatísticas em AD enjvolve construir modelos de fenômenos de AD que possuam alguma forma implícita de estrutura de dimensão mais baixa e, então, estudar os ganhos estatísticos e computacionais explorando essa estrutura.
- Para os três casos apresentados, é possível verificar como o comportamento dos procedimentos pode mudar dramaticamente quando uma estrutura de dimensão mais baixa estiver presente.
- É o que comentaremos a seguir.

Classificação: esparsidade dos vetores médias

- No exemplo de classificação, tínhamos $\hat{\mu}_j$, a média amostral de n_j amostras. Neste caso, $||\hat{\mu}_j - \mu_j||_2^2$ era concentrada ao redor de $d/n_j = \alpha$. Se $\alpha > 0$, há um nível constante de erro, de modo que o preditor clássico da taxa de erro é muito otimista.
- Quando o verdadeiro vetor de médias possui alguma estrutura de baixa dimensão, pode haver estimadores melhores.
- Talvez a forma mais simples de estrutura seja a esparsidade.
- Suponha que saibamos que cada vetor de médias μ_j seja relativamente esparso, com somente s de suas d componentes não nulas, com $s \leq d$.
- Neste caso, podemos obter um estimador substancialmente melhor aplicando-se alguma forma de limiar (threshold) às médias amostrais.
- Duas formas de limiares bastante usadas são: Hard e Soft.

Classificação: esparsidade dos vetores médias

- Hard Threshold (limiar duro): para dado $\lambda > 0$,

$$H_\lambda(x) = I\{|x| > \lambda\} = \begin{cases} x, & \text{se } x > \lambda, \\ 0, & \text{caso contrário.} \end{cases} \tag{14}$$

- Esta função é do tipo *keep-or-kill*.
- Soft Threshold limiar suave:

$$T_\lambda(x) = I\{|x > \lambda\}(x - \lambda \text{sign}(x) = \begin{cases} x - \lambda \text{sign}(x), & \text{se } |x| > \lambda \\ 0, & \text{caso contrário.} \end{cases} \tag{15}$$

- Aqui, sign(x) é igual a 1, se $x > 0$, -1, se $x < 0$ e 0 se $x = 0$.

Classificação: esparsidade dos vetores médias

- Há várias maneiras de se escolher λ. Para detalhes, veja Morettin (2014).
- Uma possibilidade é o chamado limiar universal, introduzido por Donoho e Johnstone (1994):

$$\lambda_n = \sqrt{\frac{2 \log d}{n}}. \tag{16}$$

- Há uma justificativa para essa escolha. Veja a referência acima.
- Usando o limiar duro, teremos., em nosso caso, o estimador

$$\tilde{\mu}_j = H_{\lambda_n}(\hat{\mu}_j), \;\; j = 1, 2, \tag{17}$$

com λ_n dado por (16).

Morettin, P. A. (2014). *Ondas e Ondaletas*. 2a. edição, EDUSP.

Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Classificação: esparsidade dos vetores médias

- Usando o estimador (17), podemos implementar um classificador baseado na função discriminante linear

$$\tilde{\Psi}(\mathbf{x}) = \left\langle \hat{\mu}_1 - \hat{\mu}_2, \mathbf{x} - \frac{\hat{\mu}_+ \hat{\mu}_2}{2} \right\rangle .$$

- Na Figura 4 (a)temos o gráfico dos limiares duro e suave e, na Figura 4 (b), o gráfico do erro de $\tilde{\Psi}(\mathbf{x})$ como uma função de γ (veja a equação 2).
- Em contraste com a Figura 1, o preditor clássico agora dá um ajuste excelente. Aqui, na simulação usamos os mesmos parâmetros daqueles da Figura 1: $d = 400, n = n_1 = n_2 = 800$ e $s = 5$ componentes não nulas.
- Pode-se provar que o a previsão limite clássica é exata quando $\log \frac{\binom{d}{s}}{n} \approx 0$.

Classificação: esparsidade

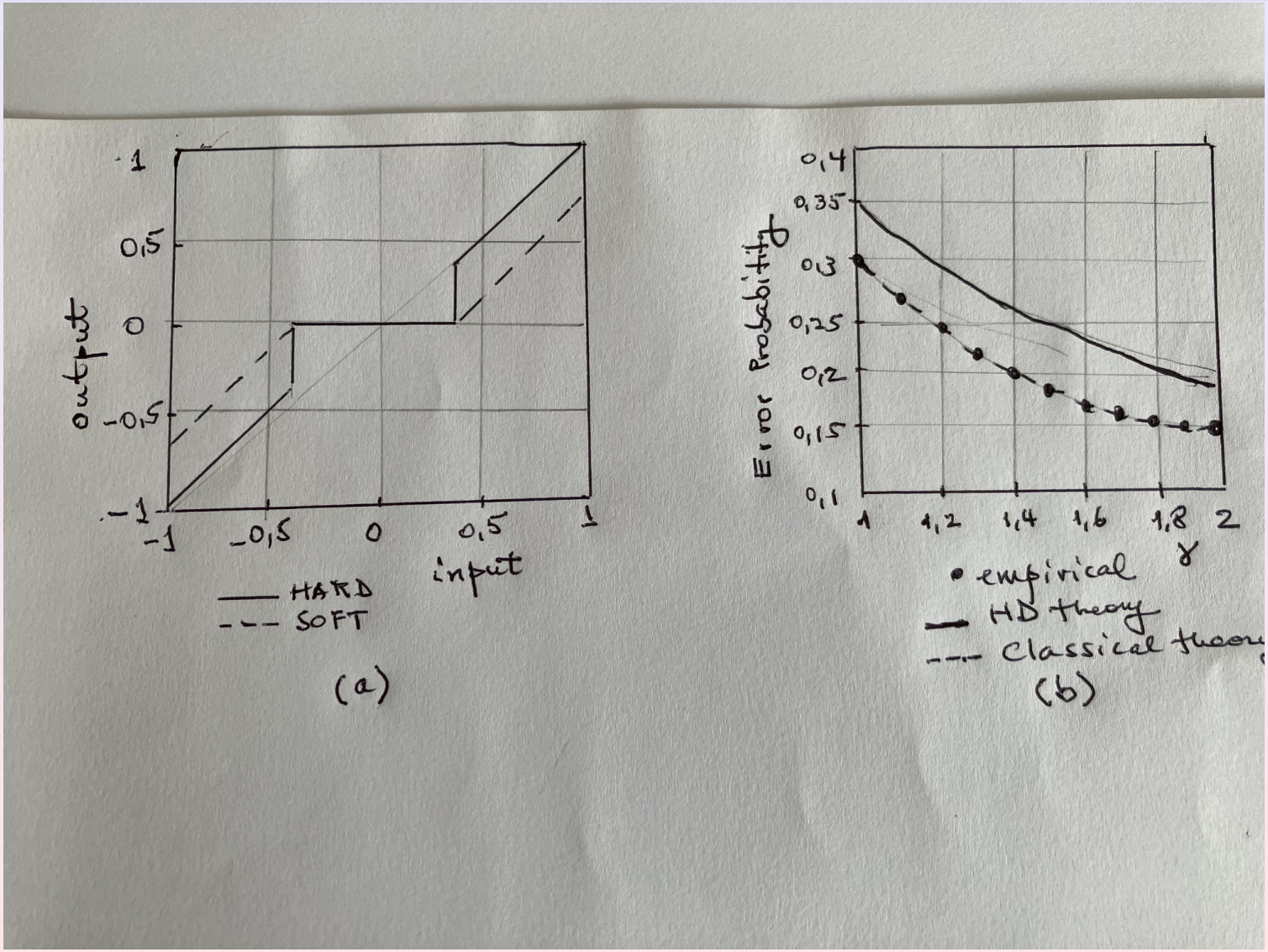


Figura 4: (a) limiares; (b) taxa de erro.

Estrutura em matrizes de covariâncias

- A questão é quais tipos de estrutura de baixa dimensão poderiam ser apropriadas para modelar matrizes de covariância.
- Um exemplo simples é supor que Σ seja diagonal. Neste caso, podemos zerar os elementos fora da diagonal principal de $\hat{\Sigma}$.
- Outra suposição mais realista é supor que Σ seja esparsa, com posições desconhecidas. Nesse caso, poderemos usar a versão limiarizada do estimador

$$\tilde{\Sigma} = T_{\lambda_n}(\hat{\Sigma}), \quad \text{com } \lambda_n = \sqrt{\frac{2 \log d}{n}}. \tag{18}$$

- A Figura 5 (a) mostra o auto-espectro $\gamma(\hat{\Sigma})$ resultante deste procedimento com $\alpha = 0,2$ e $(n, d) = (4000, 800)$, como na Figura 2.
- Em contraste com com o comportamento M-P daquela figura, agora vemos que $\gamma(\hat{\Sigma})$ está concentrado ao redor do ponto 1.
- Pode ser provado que $||\hat{\Sigma} - \Sigma||_2^2 \approx \sqrt{\frac{\log d}{n}}$ com alta probabilidade.

Estrutura em matrizes de covariâncias

- Uma forma alternativa de estrutura de baixa dimensão para matrizes **simétricas** é que seu auto-espectro decaia de maneira rápida. Considere novamente sequências de problemas indexados por (n, d) e suponha que essa sequência tenha traços limitados, ou seja, traço $(\Sigma) \leq R$, independente da dimensão d.
- Este requisito significa que os valores ordenados dos autovalores $\gamma_j(\Sigma)$ devem decair mais rápido do que j^{-1}. Esse tipo de decaimento vale em várias aplicações.
- Na Figura 5 (b) temos o log-log plot do erro $||\hat{\Sigma} - \Sigma||_2$ sobre pares (n, d), todos com razaão fixa $d/n = 0,2$, com os traços das matrizes limitados por $R = 20$.
- Resultados teóricos mostram que, para tal sequência de matrizes, o erro $||\hat{\Sigma} - \Sigma||_2$ decai como $n^{-1/2}$, mesmo que d seja da mesma ordem de n.

Estrutura em matrizes de covariâncias

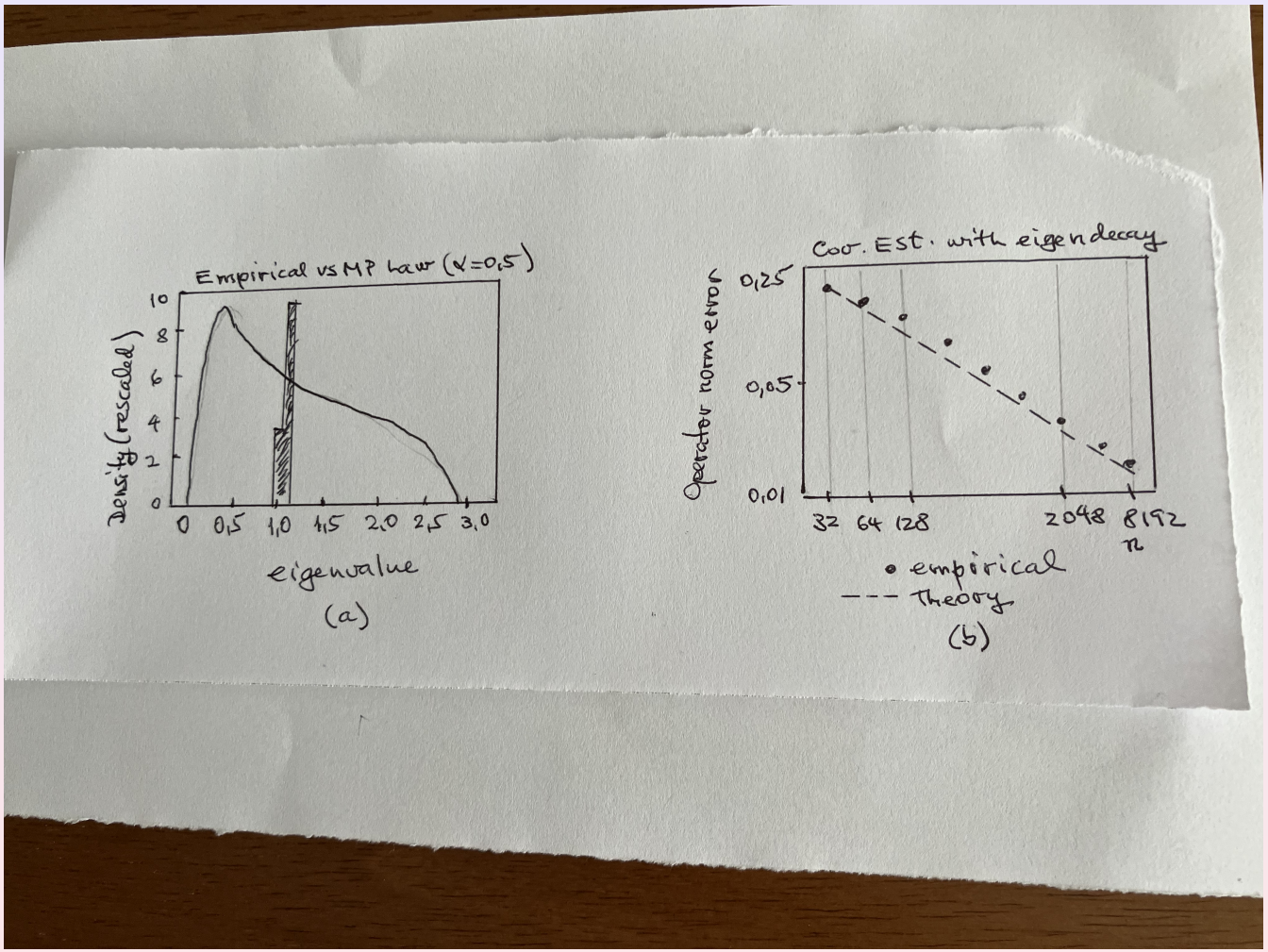


Figura 5: (a) Dados empíricos versus lei MP ($\alpha = 0,5$);
(b) Estimação da covariância com decaimentos dos autovalores.

Estruturas em regressão

- Uma forma de estrutura é aquela de uma decomposição aditiva da função:

$$f(x_1, \ldots, x_d) = \sum_{j=1}^{d} g_j(x_j), \tag{19}$$

  na qual $g_j : \mathbb{R} \to \mathbb{R}$ é escolhida de alguma classe de funções, de modo que o problema se reduz a estimar uma coleção de d funções univariadas.
- Uma forma especial de (19) é o modelo clássico linear, no qual $g_j(x_j) = \theta_j x_j$, para algum coeficiente $\theta_j \in \mathbb{R}$.
- Podemos supor, também, alguma forma de esparsidade. O modelo aditivo esparso (*sparse additive model*, SPAM) baseia-se na existência de algum subconjunto $S \subset \{1, 2, \ldots, d\}$, com cardinalidade $s = |S|$, tal que

$$f(x_1, \ldots, x_d) = \sum_{j \in S} g_j(x_j). \tag{20}$$

- Nesse caso, temos que estimar S, que pode ser um de $\binom{d}{s}$ subconjuntos possíveis de tamanho s e as funções $g_j, j \in S$.

Estruturas em regressão

- Outra forma de estrutura é o Multiple-index Model, no qual a função de regressão toma a forma

$$f(x_1, \ldots, x_d) = h(\mathbf{A}\mathbf{x}), \tag{21}$$

para alguma matriz $\mathbf{A} \in \mathbb{R}^{s \times d}$ e função $h : \mathbb{R}^s \to \mathbb{R}$.

- O Single-index Model é um caso especial com $s = 1$, de modo que $f(\mathbf{x}) = h(< \mathbf{a}, \mathbf{x} >)$, para algum vetor $\mathbf{a} \in \mathbb{R}^d$.
- Tomando somas de Single-index Models conduz a um método chamado Projection Pursuit Regression, envolvendo funções da forma

$$f(x_1, \ldots, x_d) = \sum_{j=1}^{M} g_j(< \mathbf{a}_j, \mathbf{x} >), \tag{22}$$

para alguma coleção de funções univariadas $g_j, j = 1, \ldots, M$ e uma coleção de d vetores $\mathbf{a}_j, j = 1, \ldots, M$.

Ponto de vista não-assintótico

- Podemos distinguir entre pelo menos três tipos de análise estatística, dependendo como o tamanho amostral comporta-se relativamente à dimensão e outros parâmetros.

(i) Teoria assintótica clássica. O tamanho amostral $n \to \infty$, com a dimensão d e outros parâmetros permanecendo fixos.

(ii) Teoria assintótica de alta dimensão. O par (n, d) tende ao infinito simultaneamente, enquanto para alguma função escala Ψ, a sequência $\Psi(n, d)$ permance fixa ou converge para algum valor $\alpha > 0$. Vimos exemplos em que $\Psi(n, d) = d/n$. Ψ pode depender de outros parâmetros, como no caso de esparsidade: $\Psi(n, d, s) = \log(\binom{d}{s})/n$.

(iii) Limites não assintóticos. O par (n, d) bem como outros parâmetros são vistos como fixos e afirmações com alta probabilidade são feitas como funções deles. Por exemplo como vimos em limites como (9).

- Limites desse tipo são chamados de tail bounds e concentration inequalities.
- Esses tipos de análise são relacionados. Por exemplo, os limites mencionados acima podem ser usados em provas de resultados assintóticos clássicos, como convergência (quase certa) de uma sequência de v.a.'s.