Teorie náhodných matic

Skripta: Introduction to Random Matrices: Theory and Practice
... pdf online

Co jsou náhodné matice?

- Hod kostkou ... (složitý) deterministický proces
  Pro řadu otázek nepotřebujeme znát přesné odpovědi o posloupnosti hodů.
  Kolikrát padne 1, hážeme-li 1000x ... ca. $\frac{1000}{6}$x.
  Náhodný model ("výsledkem hodu kostkou je jedno z čísel $\{1,...,6\}$ s pravd. $1/6$ každé")
  nedává odpovědi na všechny otázky ("Padne příštím hodem,
- je-li výsledkem procesu matice místo skaláru a je-li tento proces také deterministický, ale modelování (= výpočtení) náročný, pak lze výsledek modelovat náhodnými maticemi
- Hamiltonián jádra těžkých kovů lze modelovat maticí:
  Jeho přesná podoba závisí na poloze všech protonů/neutronů.
  Udat Hamiltonián jednoho konkrétního atomu je laboratorně obtížné, natož pro $\sim 10^{23}$ atomů ... je ale možné uvažovat "průměrné jádro" a "průměrný Hamiltonián"
  ... potom nás zajímají odpovídající energie ... tedy vlastní čísla

• Netflix je firma prodávající filmy online. Zákazníci hodnotí shlédnuté filmy:

	StarWars I	StarWars II	......				
Alice	2	x	x	x	4	x	...
Bob	x	x	1	x	3	2	...
Cyril	4	1	2	x	x	x	...
⋮							

Úplná matice hodnocení (N_{full}) je

• neznámá (ne všichni ohodnotí vše)

• důležitá (umožňuje předpovědět, který film se bude líbit Bobovi, ...)

• má (přibližně) nízkou hodnost

Skutečná (= pozorovaná) matice je jen projekce:

$$N_{obs} = P \cdot N_{full}$$

násobení po složkách

P má hodnoty $P_{ij} = \begin{cases} 1 \text{ ... i-tý divák hodnotil j-tý film} \\ 0 \text{ ... ne ...} \end{cases}$

Jako (nedokonalý) model analýzy databáze Netflixu

bereme $P_{ij} = 1$ s pravd. $p \ll 1$, jinak 0 ... P_{ij} nezávislé.

Náhodná matice je náhodná proměnná s oborem hodnot $\mathbb{R}^{m \times n}$ ...

- nezávislé prvky
- nezávislé řádky / sloupce
- nezávislé komponenty singulárního rozkladu
- symetrické

⋮

- Zlaté pravidlo: Máme-li proces, který generuje matice, je ale příliš složitý, můžeme zkusit uvažovat "průměrný výsledek".
- Poté nás může zajímat např.: největší / nejmenší vlastní číslo, různé normy matice, ...

Příklad náhodní matice

- $H = \begin{pmatrix} 3{,}217 & 0 & 0 \\ 0 & -2{,}31 & 0 \\ 0 & 0 & 0{,}412 \end{pmatrix}$

… diagonální matice, prvky nezávislé náhodné veličiny

… matici lze reprezentovat pomocí náhodné diagonály

… $\mathrm{diag}(H) = (3{,}217; -2{,}31; 0{,}412)$, což je i spektrum H

… vše se redukuje na studium náhodných vektorů s nezávislými komponentami

Tedy např. $\|H\|_\infty = \|H\|$ … spektrální norma H

$$= \max_{i=1,\dots,N} |H_{ii}|$$

Pokud je $H_{ii} = \omega_i \sim \mathcal{N}(0;1)$ … kolik je $\mathbb{E} \max_{i=1,\dots,N} |\omega_i|$

Cvičení: • Pro $N \in \mathbb{N}$ určete (co nejpřesněji)

$$\mathbb{E} \max_{i=1,\dots,N} \omega_i \quad \text{a} \quad \mathbb{E} \max_{i=1,\dots,N} |\omega_i|$$

- Zkuste experimentálně v Matlabu.

- $H = \begin{pmatrix} 1{,}21 & -1{,}32 & \cdot \\ -0{,}62 & 2{,}41 & \vdots \\ -3{,}2 & \cdots & \end{pmatrix}$ ... prvky jsou nezávislé $N(0;1)$ elementy

... obecně není symetrická

... my budeme studovat především reálná vlastní čísla (třebas i komplexních matic)

$H_s := \frac{H + H^T}{2}$ ... symetrická

"GOE" Gaussian orthogonal ensemble

... náhodná matice generovaná pomocí

$H_{ij} \sim N(0;1)$, $H_{ii} \sim N(0,2)$ ... $= \sqrt{2}\, N(0;1)$

$i \neq j$, nezávislé

pokud H má roz. $N(0;1)$ elementy, pak $\frac{H+H^T}{\sqrt{2}}$ je z GOE

Jednotlivé nagenerované matice nazýváme vzorky nebo instance.

Experimentálně (MATLAB) lze vidět rozložení ~~náhod~~ vlastních čísel GOE ... Fig 1.1 ... normalizovaný histogram

- Pro $N = 8$ a $T = 50000$ instancí, vyrobíme 50000 vektorů vlastních čísel, dáme do vektoru $x \in \mathbb{R}^{400000}$

Plot je jeho normalizovaný histogram (50 binů)

Krátce o náhodných proměnných

pdf ... probability density function ... hustota

- $\mathbb{P}(X \in (a,b)) = \int_a^b \rho(x)\,dx$
- $\int_{\mathbb{R}} \rho(x)\,dx = 1$ ... pdf je správně normovaná

Střední hodnota ... $\mathbb{E}X = \langle X \rangle$

Rozptyl $\mathrm{Var}(X) = \langle X^2 \rangle - \langle X \rangle^2$

Distribuční funkce $F(x) := \int_{-\infty}^{x} \rho(y)\,dy$

Nezávislé proměnné ... jpdf ... joint probability density function

$$\rho(x_1, x_2) = \rho_1(x_1)\,\rho_2(x_2)$$

Marginal pdf $\rho(x_1) = \int \rho(x_1, x_2)\,dx_2$

Příklad: • Nechť $H = (H_{ij})_{i,j=1}^{N}$, $H_{ij} \sim \mathcal{N}(0;1)$ nezávislé.

Pak $$\rho(H) = \rho(H_{11}, \dots, H_{NN}) = \prod_{i,j=1}^{N} \frac{\exp(-H_{ij}^2/2)}{\sqrt{2\pi}}.$$

- Pro $H_s = \frac{H + H^T}{2}$ je $H_{ii} \sim \mathcal{N}(0;1)$; $H_{ij} \sim \frac{1}{2}\cdot\sqrt{2}\cdot\mathcal{N}(0;1) \sim \mathcal{N}(0;1/2)$

$$\rho(H_s) = \rho\left(((H_s)_{ij})_{i \le j}\right) = \prod_{i=1}^{N} \frac{\exp(-(H_s)_{ii}^2/2)}{\sqrt{2\pi}} \cdot \prod_{i<j} \frac{\exp(-(H_s)_{ij}^2)}{\sqrt{\pi}}$$

2. Kapitola

Wigner's surmise

Buď $H_s = \begin{pmatrix} x_1 & x_3 \\ x_3 & x_2 \end{pmatrix}$, kde $x_1, x_2 \sim \mathcal{N}(0,1)$, $x_3 \sim \mathcal{N}(0,1/2)$.

Nechť $\lambda_1 < \lambda_2$ jsou její (reálná) vlastní čísla. Pak

$s = \lambda_2 - \lambda_1$ je rozdíl vlastních čísel ... spektrální mezera ...?

- Jaká je pdf $p(s)$?

$$|H_s - \lambda I| = \begin{vmatrix} x_1-\lambda & x_3 \\ x_3 & x_2-\lambda \end{vmatrix} = \lambda^2 - \lambda(x_1+x_2) + (x_1x_2 - x_3^2) \overset{!}{=} 0$$

$$\lambda_{1,2} = \frac{(x_1+x_2) \pm\sqrt{(x_1+x_2)^2 - 4(x_1x_2 - x_3^2)}}{2} = \frac{(x_1+x_2) \pm\sqrt{(x_1-x_2)^2 + 4x_3^2}}{2}$$

$$\Rightarrow s = \sqrt{(x_1-x_2)^2 + 4x_3^2}$$

- $p(s)$ tedy <u>určuje hustotu</u> pravděpodobnosti, že pro x_1, x_2, x_3 platí $s = \sqrt{(x_1-x_2)^2 + 4x_3^2}$

$$p(s) = \int_{\mathbb{R}^3} \frac{e^{-\frac{1}{2}x_1^2} e^{-\frac{1}{2}x_2^2} e^{-x_3^2}}{\sqrt{2\pi}\cdot\sqrt{2\pi}\cdot\sqrt{\pi}} \cdot \delta\left(s - \sqrt{(x_1-x_2)^2 + 4x_3^2}\right) dx_1 dx_2 dx_3$$

Změna proměnných
$$\left.\begin{aligned} x_1 - x_2 &= r\cos\theta \\ 2x_3 &= r\sin\theta \\ x_1 + x_2 &= \psi \end{aligned}\right\} \Longleftrightarrow \left\{\begin{aligned} x_1 &= \frac{r\cos\theta + \psi}{2} \\ x_2 &= \frac{\psi - r\cos\theta}{2} \\ x_3 &= \frac{r\sin\theta}{2} \end{aligned}\right.$$

$$J = \det\begin{pmatrix} \frac{\partial x_1}{\partial r}, & \frac{\partial x_1}{\partial \theta}, & \frac{\partial x_1}{\psi} \\ & \vdots & \end{pmatrix} = -r/4$$

$r^2 = (x_1-x_2)^2 + 4x_3^2$; $0 \le \theta \le 2\pi$

$0 \le r \le \infty$

$-\infty \le \psi < +\infty$

$$P(s) = \frac{1}{4}\cdot\frac{1}{(2\pi)\sqrt{\pi}}\int_{-\infty}^{\infty}\int_{0}^{2\pi}\int_{0}^{\infty} \underbrace{r}_{|J|}\cdot\delta(s-r)\cdot\exp\left[-\left(\frac{r\cos\theta+\psi}{2}\right)^2\Big/2\right]\cdot\exp\left[-\left(\frac{\psi-r\cos\theta}{2}\right)^2\Big/2\right]\cdot\exp\left[-\left(\frac{r\sin\theta}{2}\right)^2\right]\,dr\,d\theta\,d\psi$$

$$= \frac{1}{8\pi^{3/2}}\int_{-\infty}^{\infty}\int_{0}^{2\pi}\int_{0}^{\infty} r\,\delta(s-r)\exp\left\{-\frac{1}{2}\left[\frac{r^2\cos^2\theta+\psi^2}{2}+\frac{r^2\sin^2\theta}{2}\right]\right\}$$

$$= \frac{1}{8\pi^{3/2}}\int_{-\infty}^{\infty} e^{-\psi^2/4}\,d\psi\int_{0}^{\infty}\int_{0}^{2\pi} r\,\delta(s-r)\exp\left[-\frac{r^2}{4}\right]d\theta\,dr$$

$$= \frac{2\pi\cdot 2\sqrt{\pi}}{8\pi^{3/2}}\int_{0}^{\infty} r\,\delta(s-r)\,e^{-r^2/4}\,dr = \frac{1}{2}\,s\,e^{-s^2/4}.$$

Poznámky

- Poměr rozptylů na a mimo diagonálu byl zásadní
- Při stejném poměru ale jiných hodnotách
  $\Rightarrow$ přeškálování
- funkce $s \to s e^{-s^2}$ je dobrým modelem pro kompromis mezi přitahováním a odpuzováním ... Autobusy v Mexiku ...

Použití Diracovy δ-funkce vypadá záhadně...

$$p(s) = \lim_{h\to 0^+} \frac{1}{2h} \int_{s-h}^{s+h} p(t)\,dt$$

$$= \lim_{h\to 0^+} \frac{1}{2h} \mathbb{P}\Big(\overbrace{s-h \le \underbrace{\sqrt{(X_1-X_2)^2+4X_3^2}}_{r} \le s+h}^{A_{s,h}}\Big)$$

$$= \lim_{h\to 0^+} \frac{1}{2h} \int_{\mathbb{R}^3} \frac{e^{-\frac{1}{2}x_1^2-\frac{1}{2}x_2^2-x_3^2}}{\sqrt{2\pi}\sqrt{2\pi}\sqrt{\pi}} \chi_{A_{s,h}}(x_1,x_2,x_3)\,dx_1dx_2dx_3$$

$$\overset{\text{subst.}}{=} \lim_{h\to 0^+} \frac{1}{2h} \int_{-\infty}^{\infty} d\varphi \int_0^{2\pi} d\theta \int_0^{\infty} dr \left[\frac{r}{4}\cdot e^{-\varphi^2/4} e^{-r^2/4} \chi_{\{s-h\le r\le s+h\}}(r,\theta,\varphi)\right]\cdot\frac{1}{2\pi\sqrt{\pi}}$$

$$= \lim_{h\to 0^+} \frac{1}{2h}\cdot\frac{1}{8\pi^{3/2}}\cdot 2\pi \underbrace{\int_{-\infty}^{\infty} e^{-\varphi^2/4}\,d\varphi}_{2\sqrt{\pi}} \cdot \int_0^{\infty} r\cdot\chi_{\{s-h\le r\le s+h\}}(r)\, e^{-r^2/4}\,dr$$

$$= \frac{1}{2} \lim_{h\to 0^+} \frac{1}{2h} \int_0^{\infty} r\,\chi_{\{s-h\le r\le s+h\}}(r)\, e^{-r^2/4}\,dr$$

$$= \frac{1}{2} \lim_{h\to 0^+} \frac{1}{2h} \int_{s-h}^{s+h} r e^{-r^2/4}\,dr = \frac{1}{2} s e^{-s^2/4}.$$

...Delta funkce umožňuje prostší a přehlednější zápis.

Je-li H (symetrická) náhodná matice, pak její vlastní čísla jsou funkcí H, a jsou tedy rovněž náhodnými proměnnými.

... označíme-li je $x_1, \dots, x_N$, pak nás zajímá jejich rozdělení

... jpdf $\rho(x_1, \dots, x_N)$

Nejjednodušší situace: diagonální matice

- $H = \begin{pmatrix} x_1 & & 0 \\ & \ddots & \\ 0 & & x_N \end{pmatrix}$, $x_1, \dots, x_N$ nezávislé a stejné rozdělení

  pak $x_1, \dots, x_N$ jsou i vlastní čísla H.

  $\rho(x_1, \dots, x_N) = \rho(x_1) \dots \rho(x_N)$

- Jak je to s mezerami mezi $x_1, \dots, x_N$? Je zde nějaká repulze?

Buď $X_1, \dots, X_N$ nezávislé náhodné proměnné, hustota $P_X(x)$, distr. funkce $F(x)$

- $P_N(s \mid X_j = x)$ ... hustota toho, že pokud je $X_j = x$, pak existuje X_k, $k \neq j$ s $X_k = s + x$ a žádná proměnná mezi tím

- Tvrdíme, že $P_N(s \mid X_j = x) = P_X(x+s)\left[F(x) + 1 - F(x+s)\right]^{N-2} \cdot \underline{(N-1)}$

  extra k (2.6)

Zdůvodnění ... $X_j = x$, nějaká ze zbývajících $N-1$ proměnných je (okolo) $x+s$, ostatních $(N-2)$ v intervalu $(-\infty, x) \cup (x+s, \infty)$.

- Kontrola normalizace

Pro $X_j = x$ se může stát, že existuje jiná $X_k = x+s$, $s>0$

nebo $X_j = x$ je největší ze všech $X_1, \dots, X_N$

Mělo by tedy být

$$\int_0^\infty p_N(s|X_j=x)\,ds + F(x)^{N-1} = 1 \ ?!$$

Opravdu: $$\int_0^\infty p_X(x+s)\left[1-F(x+s)+F(x)\right]^{N-2} ds\cdot(N-1) + F(x)^{N-1}$$

$u = F(x+s)$, $u' = F'(x+s) = p_X(x+s)$

$$= \underbrace{\int_{F(x)}^1 \left[1+F(x)-u\right]^{N-2} du}\cdot(N-1) + F(x)^{N-1}$$

$$= \frac{1-F(x)^{N-1}}{N-1}\cdot(N-1) + F(x)^{N-1} = 1.$$

- Hustota jevu mezery s při libovolné proměnné rovné x:

$$p_N(s \,\&\, \text{any } X=x) = \sum_{j=1}^N p_N(s \,\&\, X_j = x) = \text{"}\sum_{j=1}^N \frac{p_N(s \,\&\, X_j=x)}{\mathbb{P}(X_j=x)}\cdot\mathbb{P}(X_j=x)\text{"}$$

$$= \sum_{j=1}^N p_N(s|X_j=x)\cdot p_X(x) = N\,p_N(s|X_j=x)\,p_X(x)$$

- Pravděpodobnost, že mezi $X_1, \dots, X_N$ je někde mezera s, nezávisle na x:

$$p_N(s) = \int_{-\infty}^\infty p_N(s \,\&\, \text{any } X=x)\,dx = N \int_{-\infty}^\infty p_X(x)\, p_N(s|X_j=x)\,dx$$

- Opět kontrola normalizace ... $\int_0^\infty p_N(s)\,ds = \underset{\uparrow}{N-1}$
  počet mezer mezi $X_1, \dots, X_N$

$$\int_0^\infty p_N(s)\,ds = \int_0^\infty N \int_{-\infty}^\infty p_X(x)\, p_N(s \mid X_j = x)\,dx\,ds$$

$$= N \int_0^\infty \int_{-\infty}^\infty p_X(x) \cdot (N-1)\, p_X(x+s)\left[F(x)+1-F(x+s)\right]^{N-2} dx\,ds$$

$$= N\cdot(N-1) \int_{-\infty}^\infty p_X(x) \int_0^\infty p_X(x+s)\left[F(x)+1-F(x+s)\right]^{N-2} ds\,dx \qquad u = F(x+s)$$

$$= N(N-1)\int_{-\infty}^\infty p_X(x) \int_{F(x)}^1 \left[F(x)+1-u\right]^{N-2} du\,dx$$

$$= N(N-1)\int_{-\infty}^\infty p_X(x)\, \frac{1-F(x)^{N-1}}{N-1}\,dx = N\cdot \int_0^1 \left[1-v^{N-1}\right]dv = N - N\cdot\frac{1}{N} = N-1.$$

$$F(x) = v$$
$$p_X(x) = F'(x) = dv$$

- Zkusme uvažovat $\lim_{N\to+\infty}$

Pokud např. $X_1, \dots, X_N \in [0;1]$ rovnoměrně rozdělené,
pak vzdálenosti mezi nimi samozřejmě klesají
s $N\to+\infty$ ... více hodnot ve stejném intervalu ... $1/N$
Místo s budeme tedy uvažovat $N\cdot s$, přesněji

$$\hat{s} = s\cdot N\cdot p_X(x)$$

• Pro jednoduchost $p_X(x) = 1$ na $[0;1]$

[Substituce v hustotě .... je-li $\int_{\mathbb{R}} p(\tau)\,d\tau = 1$, pak

$$\tilde{p}(\mu) = a \cdot p(a\mu),\ a > 0 \text{ je opět hustota}$$

$$\int_{\mathbb{R}} \tilde{p}(\mu)\,d\mu = \int_{\mathbb{R}} a\, p(a\mu)\,d\mu = \int_{\mathbb{R}} p(\tau)\,d\tau = 1$$

$$\tau = a\mu$$
$$d\tau = a\,d\mu$$

Např. $a = 1/2$ $\quad \tilde{p}(\mu) = \frac{1}{2}\chi_{[0,2]}(\mu)$.

• $$\frac{1}{N} p_X\left(s = \frac{\hat{s}}{N} \,\Big|\, X_j = x\right) = \frac{N-1}{N} p_X\left(x + \frac{\hat{s}}{N}\right)\left[1 + F(x) - F\left(x + \frac{\hat{s}}{N}\right)\right]^{N-2}$$

Pro $N \to +\infty$, $\hat{s} \sim 1$ je

$$p_X\left(x + \frac{\hat{s}}{N}\right) \to p_X(x)$$
$$F\left(x + \frac{\hat{s}}{N}\right) - F(x) \sim F'(x)\cdot\frac{\hat{s}}{N} = \underbrace{p_X(x)}_{=1}\cdot\frac{\hat{s}}{N} = \frac{\hat{s}}{N}$$

$$\to 1 \cdot \overbrace{p_X(x)}^{=1} \cdot \left[1 - \frac{\hat{s}}{N}\right]^{N-2} \to e^{-\hat{s}}.$$

Ať je tedy $X_j = x$ jakékoliv, malé mezery k dalším X_k nejsou méně pravděpodobné než mezery větší!

- Jpdf pro Gaussovské matice (GOE)

Jpdf vlastních čísel $N\times N$ Gaussovské matice je

$$\varrho(x_1,\dots,x_N) = \frac{1}{Z_{N,\beta}}\, e^{-\frac{1}{2}(x_1^2+\dots+x_N^2)} \prod_{j<k} |x_j - x_k|^\beta ,$$

$$Z_{N,\beta} = (2\pi)^{N/2} \prod_{j=1}^{N} \frac{\Gamma(1+j\beta/2)}{\Gamma(1+\beta/2)}$$ je normalizační konstanta

- $\beta = 1$ pro GOE
- vlastní čísla jsou v této reprezentaci neuspořádaná

Důkaz bude později, až zavedeme několik jedn. vlastností

- Gaussovský faktor „eliminuje" konfigurace, kde $|x_1|,\dots,|x_N|$ jsou velké
- Faktor $\prod_{j<k}|x_j - x_k|$ znevýhodňuje vlastní čísla stejné velikosti
- ⇓ není faktorizovatelný ... každé x_j není nezávislé na všech ostatních x_k, $k \neq j$.

Cvičení: Wigner's surmise je důsledek pro $N=2$

$$\dots\ p(s) = \int_{\mathbb{R}^2} \varrho(x_1,x_2)\,\delta(s - |x_1 - x_2|)\, d(x_1,x_2)\ \dots\ ?$$

$$\iint_{-\infty}^{\infty} dx_1\,dx_2\; \rho(x_1,x_2)\,\delta(s-|x_2-x_1|)$$

$$= \iint_{-\infty}^{\infty} dx_1\,dx_2\; \frac{1}{Z_{2,1}}\, e^{-\frac{1}{2}(x_1^2+x_2^2)}\,|x_1-x_2|\,\delta(s-|x_2-x_1|)$$

$$Z_{2,1} = (2\pi)\,\frac{\Gamma(1+1/2)}{\Gamma(1+1/2)}\cdot\frac{\Gamma(1+1)}{\Gamma(1+1/2)} = 2\pi\cdot\frac{1}{\frac{1}{2}\sqrt{\pi}} = \frac{4\sqrt{\pi}}{\cancel{\sqrt{2\pi}}}$$

$$= \frac{\cancel{\sqrt{2\pi}}}{4\sqrt{\pi}}\iint_{-\infty}^{\infty} dx_1\,dx_2\; e^{-\frac{1}{2}(x_1^2+x_2^2)}\,|x_1-x_2|\,\delta(s-|x_2-x_1|)$$

$$u = x_1-x_2 \qquad x_1 = \frac{u+v}{2},\; x_2 = \frac{v-u}{2}$$

$$v = x_1+x_2$$

$$x_1^2+x_2^2 = \frac{u^2+2uv+v^2}{4} + \frac{v^2-2uv+u^2}{4} = \frac{u^2+v^2}{2}$$

$$\begin{pmatrix} u \\ v \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}}_{\det = 2}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$= \frac{\cancel{\sqrt{2\pi}}}{4\sqrt{\pi}}\iint_{\mathbb{R}} e^{-\frac{(u^2+v^2)}{4}}\,|u|\,\delta(s-|u|)\,du\,dv\cdot 2^{-1}$$

$$= \frac{\cancel{\sqrt{2\pi}}}{8\sqrt{\pi}}\cdot 2\cdot\int_0^{\infty} du \int_{-\infty}^{\infty} dv\; e^{-\frac{u^2+v^2}{4}}\cdot u\,\delta(s-u) = \frac{1}{4\sqrt{\pi}}\int_0^{\infty} u\,\delta(s-u)\,e^{-u^2/4}\,du\,\underbrace{\int_{-\infty}^{\infty} e^{-v^2/4}\,dv}_{2\sqrt{\pi}}$$

$$= 2^{-1}\int_0^{\infty} u\,\delta(s-u)\,e^{-u^2/4}\,du = 2^{-1}\,s\,e^{-s^2/4}$$