

# Zmiešavanie diskretných distribúcií

## Dizertačná práca

Samuel Hudec

Univerzita Mateja Bela  
Fakulta prírodných vied  
Katedra matematiky

22. 8. 2019

# Ciele prezentacie

- Priemerované zmiešané distribúcie
- V dizertačnej práci zahrnuté priemerované zmiešané distribúcie
- Parametrické priestory
- Charakteristiky distribúcie
- Odhadovanie parametrov
- Testy dobrej zhody
- Publikačná činnosť

# Priemerované zmiešané distribúcie

Nech  $X_1, X_2, \dots, X_n$  sú nezávislé diskkrétne náhodné premenné nadobúdajúce hodnoty  $x_1, x_2, \dots$  zo známymi (štandardnými) distribúciami pravdepodobnosti a pmf tvaru  $\{\frac{f_j(x_i)}{\alpha^{(j)}}\}_{i=1}^{\infty}$ ,  $j = 1, 2, \dots, n$ . Potom priemerovaná diskrétna zmiešaná distribúcia má pravdepodobnostnú funkciu

$$P(x_i) = C \left( \sum_{j=1}^n a_j f_j(x_i) \right), \quad i = 1, 2, \dots$$

$f_j(x_i)$  sú „nenormalizované časti“ známych distribúcií pravdepodobnosti s váhami  $a_j$  a  $C^{-1} = \sum_{j=1}^n a_j \alpha^{(j)}$  je normalizačná konštanta.

# Zahrnuté priemerované zmiešané distribúcie

- Averaged mixed logarithmic distribution Type 1a
- Averaged mixed logarithmic distribution Type 2

# Zahrnuté priemerované zmiešané distribúcie

- Averaged mixed logarithmic distribution Type 1a
- Averaged mixed logarithmic distribution Type 2
- Averaged mixed logarithmic-geometric distribution Type 3a
- Averaged mixed logarithmic-geometric distribution Type 3b
- Averaged mixed logarithmic-geometric distribution Type 3c

# Zahrnuté priemerované zmiešané distribúcie

- Averaged mixed logarithmic distribution Type 1a
- Averaged mixed logarithmic distribution Type 2
- Averaged mixed logarithmic-geometric distribution Type 3a
- Averaged mixed logarithmic-geometric distribution Type 3b
- Averaged mixed logarithmic-geometric distribution Type 3c
- Averaged mixed logarithmic-negative binomial distribution

# Zahrnuté priemerované zmiešané distribúcie

## Averaged mixed logarithmic distribution Type 1a

$$P(x) = \frac{a^x - rb^x}{x[r \ln(1-b) - \ln(1-a)]}, \quad \text{pre } x = 1, 2, 3, \dots$$

## Averaged mixed logarithmic distribution Type 2

$$P(x) = \frac{a^x + (-1)^{x+1}rb^x}{x[r \ln(1+b) - \ln(1-a)]}, \quad \text{pre } x = 1, 2, 3, \dots$$

## Averaged mixed logarithmic-geometric distribution Type 3a

$$P(x) = \begin{cases} \frac{r}{\frac{r}{(1-q)} - \ln(1-\theta)}, & \text{pre } x = 0, \\ \frac{rq^x + \frac{\theta^x}{x}}{\frac{r}{(1-q)} - \ln(1-\theta)}, & \text{pre } x = 1, 2, \dots \end{cases}$$

## Averaged mixed logarithmic-geometric distribution Type 3b

$$P(x) = \frac{rq^x + \frac{\theta^x}{x}}{rq(1-q)^{-1} - \ln(1-\theta)}, \quad \text{pre } x = 1, 2, 3, \dots$$

## Averaged mixed logarithmic-geometric distribution Type 3c

$$P(x) = \begin{cases} \frac{r}{r(1-b)^{-1} + \ln(\frac{1-a}{1-b})}, & \text{pre } x = 0, \\ \frac{rb^x + \frac{b^x - a^x}{x}}{r(1-b)^{-1} + \ln(\frac{1-a}{1-b})}, & \text{pre } x = 1, 2, \dots \end{cases}$$



# Zahrnuté priemerované zmiešané distribúcie

## Averaged mixed logarithmic-negative binomial distribution

$$P(x) = \begin{cases} \frac{r(1-q)^m}{r - \ln(1-q)}, & \text{pre } x = 0, \\ \frac{r \binom{m+x-1}{x} (1-q)^m q^x + \frac{q^x}{x}}{r - \ln(1-q)}, & \text{pre } x = 1, 2, \dots \end{cases}$$

pričom

$$\binom{m+x-1}{x} = \begin{cases} \binom{m+x-1}{x} = 1, & \text{pre } x = 0, \\ \binom{m+x-1}{x} = \frac{m(m-1)(m-2)\dots(m-x+1)}{x!}, & \text{pre } x = 1, 2, \dots \end{cases}$$

# Parametrický priestor

Na základe podmienok definície pravdepodobnostnej funkcie

$$P(x_i) \geq 0, \quad i = 1, 2, 3, \dots$$

$$\sum_{i=1}^{\infty} P(x_i) = 1.$$

# Parametrický priestor

Na základe podmienok definície pravdepodobnostnej funkcie

$$P(x_i) \geq 0, \quad i = 1, 2, 3, \dots$$

$$\sum_{i=1}^{\infty} P(x_i) = 1.$$

$$\sum_{j=1}^n a_j f_j(x_i) \geq 0, \quad \text{pre } i = 1, 2, 3, \dots$$

$$\sum_{j=1}^n a_j f_j(x_i) \leq 0, \quad \text{pre } i = 1, 2, 3, \dots$$

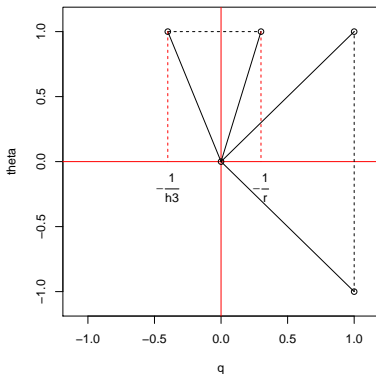
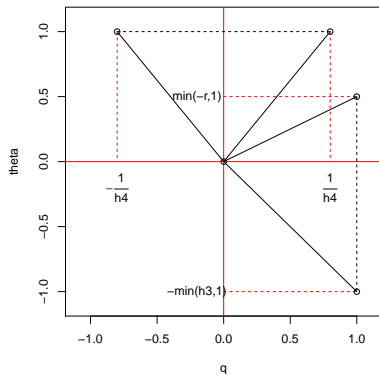
sme odvodzovali parametrický priestor každej distribúcie.

# Parametrický priestor

Ako príklad výsledku zdĺhavých odvodzovaní súhrnný parametrický priestor pre Averaged mixed logarithmic-geometric distribution Type 3b

- 1  $\{(r, q, \theta) : r \geq e, \quad \{0 \leq q < 1, -q \leq \theta < 1\} \cup \{-1 < q < 0, -qr \leq \theta < 1\}\}$
- 2  $\{(r, q, \theta) : 0 < r < e, \quad \{0 \leq q < 1, -q \min\{r, 1\} \leq \theta < 1\} \cup \{h = \max\{(2k_0+1)\sqrt{(2k_0+1)r}, (2k_0+3)\sqrt{(2k_0+3)r}\}, -1 < q < 0, -qh \leq \theta < 1\}\}$
- 3  $\{(r, q, \theta) : r = 0, \quad 0 < \theta < 1, -1 < q < 1\}$
- 4  $\{(r, q, \theta) : -e \leq r < 0, \quad \{0 \leq q < 1, \theta \in \langle -q \min\{\sqrt{2(-r)}, 1\}, q \min\{-r, 1\}\rangle\}\}$
- 5  $\{(r, q, \theta) : r \leq -e, \quad \{0 \leq q < 1, \theta \in \langle -q, q \rangle\}\}$

# Parametrický priestor



Obr.: Zložený parametrický priestor pre  $-e < r \leq 0$  (vľavo) a  $r \leq -e$  (vpravo), kde  $h3 = \sqrt{2(-r)}$  a  $h4 = \max\{^{(x_0)}\sqrt{x_0(-r)}, ^{(x_0+1)}\sqrt{(x_0+1)(-r)}\}$

# Charakteristiky distribúcie

Pokračujme pre Averaged mixed logarithmic-geometric distribution Type 3b.  
Vytvárajúca funkcia

$$G(t) = \frac{rq t(1 - qt)^{-1} - \ln(1 - \theta t)}{rq(1 - q)^{-1} - \ln(1 - \theta)}.$$

Klesajúci faktoriálny moment  $t$ -teho rádu je

$$\mu'_{[t]} = \frac{1}{rq(1 - q)^{-1} - \ln(1 - \theta)} \left[ \frac{t!rq^t}{(1 - q)^{t+1}} + \frac{(t - 1)! \theta^t}{(1 - \theta)^t} \right].$$

# Charakteristiky distribúcie

Momenty, kde označme  $C = rq(1 - q)^{-1} - \ln(1 - \theta)$

$$\mu'_1 = \frac{1}{C} \left( \frac{rq}{(1 - q)^2} + \frac{\theta}{1 - \theta} \right),$$

$$\mu'_2 = \frac{1}{C} \left( -\frac{rq(q + 1)}{(q - 1)^3} + \frac{\theta}{(1 - \theta)^2} \right),$$

$$\mu'_3 = \frac{1}{C} \left( \frac{rq(q^2 + 4q + 1)}{(q - 1)^4} - \frac{\theta(\theta + 1)}{(\theta - 1)^3} \right),$$

$$\mu'_4 = \frac{1}{C} \left( -\frac{rq(q^3 + 11q^2 + 11q + 1)}{(q - 1)^5} + \frac{\theta(\theta^2 + 4\theta + 1)}{(\theta - 1)^4} \right).$$

# Charakteristiky distribúcie

## Centrálné momenty

$$\mu_1 = 0,$$

$$\mu_2 = \frac{1}{C^2} \left( -\frac{rCq(q+1)}{(q-1)^3} + \frac{C\theta}{(1-\theta)^2} - \left( \frac{rq}{(1-q)^2} + \frac{\theta}{1-\theta} \right)^2 \right),$$

$$\mu_3 = \frac{1}{C} \left( \frac{rq(q^2 + 4q + 1)}{(q-1)^4} - \frac{\theta(\theta+1)}{(\theta-1)^3} \right) - \frac{3}{C^2} \left( -\frac{rq(q+1)}{(q-1)^3} + \frac{\theta}{(1-\theta)^2} \right) \left( \frac{rq}{(1-q)^2} + \frac{\theta}{1-\theta} \right) + \frac{2}{C^3} \left( \frac{rq}{(1-q)^2} + \frac{\theta}{1-\theta} \right)^3,$$

$$\mu_4 = \frac{1}{C} \left( -\frac{rq(q^3 + 11q^2 + 11q + 1)}{(q-1)^5} + \frac{\theta(\theta^2 + 4\theta + 1)}{(\theta-1)^4} \right) - \frac{4}{C^2} \left( \frac{rq(q^2 + 4q + 1)}{(q-1)^4} - \frac{\theta(\theta+1)}{(\theta-1)^3} \right) \left( \frac{rq}{(1-q)^2} + \frac{\theta}{1-\theta} \right) + \frac{6}{C^3} \left( -\frac{rq(q+1)}{(q-1)^3} + \frac{\theta}{(1-\theta)^2} \right) \left( \frac{rq}{(1-q)^2} + \frac{\theta}{1-\theta} \right)^2 - \frac{3}{C^4} \left( \frac{rq}{(1-q)^2} + \frac{\theta}{1-\theta} \right)^4,$$

...



# Odhady parametrov

- Metóda štyroch frekvencií
- Momentova metóda
- Kombinácie

# Odhady parametrov

- Metóda štyroch frekvencií
- Momentová metóda
- Kombinácie
- Metóda maximálnej vierohodnosti
- Metóda minimálneho  $\chi^2$
- Iteračne Metóda maximálnej vierohodnosti
- Iteračne Metóda minimálneho  $\chi^2$

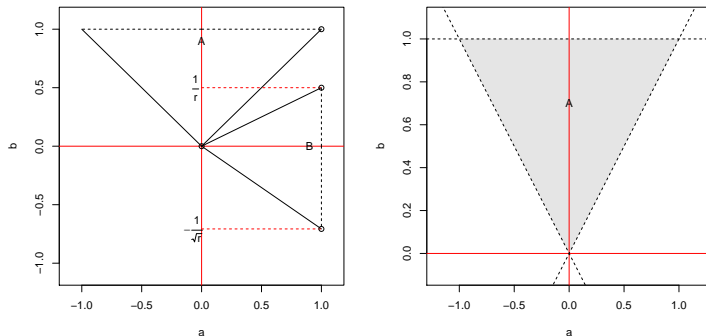
- log-likelihood (logaritmická vierohodnostná) funkcia

$$\ln \mathcal{L}(x_1, x_2, \dots, x_n; \theta) = \sum_{i=1}^n \ln P(x_i).$$

- Funkciu použijeme na numerické optimalizovanie za nelineárnych reštrikcii (parametrický priestor).
- `constrOptim.nl` z knižnice `alabama`, ktorá využíva `Augmented Lagrangian Adaptive Barrier Minimization Algorithm`.

# Odhady parametrov

Averaged mixed logarithmic distribution Type 1a



Obr.: Parametrický priestor pre  $r > 1$  (vľavo). Priamky tvoriace hranice časti A parametrického priestoru vo forme  $h(x) > 0$  (vpravo).

# Odhady parametrov

- 1 Algoritmus necháme iterovať na týchto oblastiach a nájdeme hodnoty  $\theta$  v ktorých  $\ln \mathcal{L}(x_1, x_2, \dots, x_n; \theta)$  nadobúda maximum,
- 2 porovnáme tieto odhady a vyberieme ten, v ktorom je hodnota funkcie  $\ln \mathcal{L}(x_1, x_2, \dots, x_n; \theta)$  najvyššia.

# Odhady parametrov

- 1 Algoritmus necháme iterovať na týchto oblastiach a nájdeme hodnoty  $\theta$  v ktorých  $\ln \mathcal{L}(x_1, x_2, \dots, x_n; \theta)$  nadobúda maximum,
- 2 porovnáme tieto odhady a vyberieme ten, v ktorom je hodnota funkcie  $\ln \mathcal{L}(x_1, x_2, \dots, x_n; \theta)$  najvyššia.

Na otestovanie správnosti sme vzali fixné série parametrov, nagenerujeme 1000 pseudo-náhodných výberov a pre každý odhadneme parametre.

# Odhady parametrov

- 1 Algoritmus necháme iterovať na týchto oblastiach a nájdeme hodnoty  $\theta$  v ktorých  $\ln \mathcal{L}(x_1, x_2, \dots, x_n; \theta)$  nadobúda maximum,
- 2 porovnáme tieto odhady a vyberieme ten, v ktorom je hodnota funkcie  $\ln \mathcal{L}(x_1, x_2, \dots, x_n; \theta)$  najvyššia.

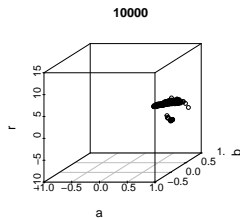
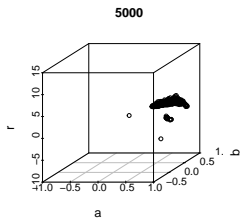
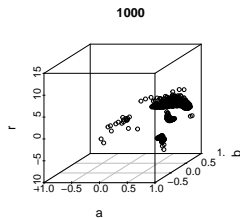
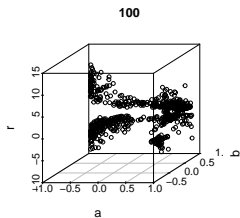
Na otestovanie správnosti sme vzali fixné série parametrov, nagenerujeme 1000 pseudo-náhodných výberov a pre každý odhadneme parametre.

Nepodarilo sa nám nájsť žiaden vhodný počiatočný odhad a preto sme zvolili pevné štartovacie hodnoty.

Rovnaký postup sme zvolili aj pri Metóda minimalneho  $\chi^2$ .

# Odhady parametrov

Averaged mixed logarithmic distribution Type 1a





# Odhady parametrov

Averaged mixed logarithmic-geometric distribution Type 3b

Metóda maximálnej vierohodnosti

#	r=4	q=0.5	$\theta=0.9$	r=2	q=0.8	$\theta=0.3$
N=50	1.1991	0.26005	0.61262	-0.66969	0.53507	0.32658
N=100	1.7044	0.28938	0.73745	-0.88347	0.58552	0.32828
N=500	1.6606	0.41995	0.83597	0.017664	0.71405	0.25674
N=1000	1.8612	0.4683	0.84959	0.052443	0.74842	0.19796
N=5000	2.2730	0.54707	0.87000	1.147175	0.79988	0.24344
N=10000	0.10427	0.88608	0.79058	1.61569	0.8004	0.29395

#	r=-2	q=0.6	$\theta=0.2$	r=-5	q=0.9	$\theta=-0.3$
N=50	-0.77586	0.27438	0.37635	-0.68115	0.74425	0.24507
N=100	-0.67194	0.32862	0.34155	-0.43597	0.81250	0.12256
N=500	-0.65935	0.39971	0.32358	-0.10142	0.88970	0.06361
N=1000	-0.5123	0.43435	0.30001	-0.26131	0.89883	0.00974
N=5000	-1.2744	0.52087	0.2772	-0.40228	0.90017	-0.02577
N=10000	-1.3444	0.56801	0.23775	-1.3947	0.90011	-0.07790

# Odhady parametrov

Averaged mixed logarithmic-geometric distribution Type 3b

Metóda minimálneho  $\chi^2$

#	r=4	q=0.5	$\theta=0.9$	r=2	q=0.8	$\theta=0.3$
N=50	-1.26742	0.26925	0.30936	-1.04517	0.44142	0.42434
N=100	-0.47203	0.18293	0.45464	-0.55110	0.64915	0.22500
N=500	0.07009	0.03766	0.78329	-0.06827	0.74813	0.14465
N=1000	0.75400	0.09731	0.81673	0.28141	0.76662	0.18890
N=5000	1.86974	0.61814	0.85049	1.43880	0.80013	0.29334
N=10000	2.29425	0.62027	0.86215	1.75130	0.80112	0.32057

#	r=-2	q=0.6	$\theta=0.2$	r=-5	q=0.9	$\theta=-0.3$
N=50	-1.20604	0.27127	0.34043	-1.61221	0.42050	0.60160
N=100	-0.18855	0.35629	0.27378	-0.77197	0.76170	0.174459
N=500	-0.42670	0.40550	0.30707	-1.08489	0.89109	-0.02281
N=1000	0.03947	0.44611	0.25683	-1.14776	0.89859	-0.02005
N=5000	-0.75157	0.50774	0.25609	-1.55814	0.90067	-0.03910
N=10000	-1.40409	0.55344	0.24074	-1.65347	0.90049	-0.07229

Averaged mixed logarithmic distribution Typu 1a má asymptotické vlastnosti odhadu maximálnej vierohodnosti

$$\hat{\boldsymbol{\theta}} \approx N_3\left(\boldsymbol{\theta}; \frac{1}{n}\mathbf{J}^{-1}(\boldsymbol{\theta})\right),$$

Napiseme 95% intervaly spoľahlivosti pre parametre  $r, a, b$

$$\begin{aligned} &\left(\hat{a} - u_{0,025}\sqrt{(1/N)\mathbf{J}_{11}^{-1}(\hat{r}, \hat{a}, \hat{b})}, \quad \hat{a} + u_{0,025}\sqrt{(1/N)\mathbf{J}_{11}^{-1}(\hat{r}, \hat{a}, \hat{b})}\right), \\ &\left(\hat{b} - u_{0,025}\sqrt{(1/N)\mathbf{J}_{22}^{-1}(\hat{r}, \hat{a}, \hat{b})}, \quad \hat{b} + u_{0,025}\sqrt{(1/N)\mathbf{J}_{22}^{-1}(\hat{r}, \hat{a}, \hat{b})}\right), \\ &\left(\hat{r} - u_{0,025}\sqrt{(1/N)\mathbf{J}_{33}^{-1}(\hat{r}, \hat{a}, \hat{b})}, \quad \hat{r} + u_{0,025}\sqrt{(1/N)\mathbf{J}_{33}^{-1}(\hat{r}, \hat{a}, \hat{b})}\right), \end{aligned}$$

# Testy dobrej zhody

Nech  $\xi_i^{(n)}$  sú empirické početnosti hodnôt z realizácie náhodného výberu  $X_1, X_2, \dots, X_n$ , ktoré patria do  $i$ -teho intervalu. Odhad metódou minimálneho  $\chi^2$  vektora parametrov budeme označovať  $\tilde{\boldsymbol{\theta}} = (\tilde{r}, \tilde{a}, \tilde{b})$ . Ak realizácia

$$\sum_{i=1}^k \frac{(\xi_i^{(n)(real)} - np_i(\tilde{\boldsymbol{\theta}})^{(real)})^2}{np_i(\tilde{\boldsymbol{\theta}})^{(real)}} \geq \chi_{k-1-u}^2(1-\alpha),$$

kde  $\chi_{k-1-u}^2(1-\alpha)$  je  $(1-\alpha)$  kvantil distribúcie  $\chi^2$ , tak zamietame hypotézu, že náhodný výber pochádza z „našej“ distribúcie na hladine významnosti  $\alpha$ .

# Testy dobrej zhody

Vysledky odhadu chyby prvého druhu pre fixné trojice parametrov pre

Averaged mixed logarithmic-geometric distribution Type 3b

#	$r \geq e$	$e > r > 0$	$-e < r < 0$	$-r \leq e$
N=50	0.969	0.735	0.884	0.974
N=100	0.625	0.104	0.334	0.189
N=500	0.1	0.065	0.074	0.085
N=1000	0.307	0.063	0.07	0.061
N=5000	0.485	0.051	0.069	0.054
N=10000	0.41	0.054	0.055	0.063

- Hudec S., *Priemerované zmiešané rozdelenia typu 1a*, Forum Statisticum Slovacum 1/2017
- Hudec S., *Priemerované diskkrétne zmiešané logaritmické rozdelenia*, ROBUST 2018
- Hudec S., *Cluster analysis on panel data*, Forum Statisticum Slovacum 1/2018
- Hudec S., *Generalizations of the lasso penalty*, WDS-m, Prague 2018
- Hudec S., Kiaba M., Knapková M., *Impact of macroeconomic indicators on public debt of Slovak Republic* Journal of Business Economics and Management, vol. 20, no. 4 (2019)

- Hudec S., Špírková J., *Smoothing of mortality rates using mixture functions*, IPMU 2018
- Hudec S., Špírková J., *Mixture function as an appropriate smoothing of mortality rates*, RELIK 2017
- Hudec S., *Modelling The Force of Mortality Using Local Polynomial Method in R*, 20th Application of Mathematics and Statistics in Economics 2017
- Hudec S., Špírková J., *Mixture function in mortality rates aggregation*, The 3rd International Symposium on Fuzzy Sets - Uncertainty Modelling 2017
- Hudec S., Gubalová J., Medved'ová P., Špírková J., *The impact of smoothing mortality rates on life insurance*, RELIK 2018





# Otázky a pripomienky oponentov

## Otázka

Ktoré najaktuálnejšie podnety z praxe viedli ku skúmaným distribúciám?

## Pripomienka

v rozdelení prametrickeho priestoru na strane 23 chyba prípad  $r = 0$  hoci sa autor v ďalšom aj týmto prípadom zaoberá.

## Pripomienka

na strane 24 v riadku 12 je uvedené "pre prípad  $r > 1$  a  $y = 1, 2, \dots$  je funkcia  $f(y)$  klesajúca", v skutočnosti je ale klesajúca iba pre  $r > 1$ , pretože pre každé konkrétne  $y$  je  $f(y)$  reálne číslo a nie funkcia. Hneď na ďalšej strane v prvom riadku sa hovorí o rade, pričom ale ide o postupnosť - rad evokuje súvislosť s nekonečným radom, o ktorý tu ale nejde.

## Pripomienka

Rovnako pri prípadnej publikácii výsledkov odporúčam vyhýbať sa hovorovému štýlu (napr. "limitne ide" má byť "jej limita je")



# Otázky a pripomienky oponentov

Median

#	$r=4$	$q=0.5$	$\theta=0.9$
N=50	0.08566229	0.2477554	0.7505398
N=100	0.2003034	0.3637268	0.7876622
N=500	0.8624809	0.4924508	0.8180208
N=1000	2.66106	0.4906059	0.8742398
N=5000	2.718282	0.4960561	0.8853658
N=10000	0.006636402	0.9032513	0.7881391
N=20000	0.007502586	0.8985859	0.7917735
N=50000	0.006516578	0.9033194	0.7879641

# Otázky a pripomienky oponentov

Mean

#	$r=4$	$q=0.5$	$\theta=0.9$
N=50	0.6270069	0.2192697	0.6078167
N=100	1.352762	0.2462451	0.7259171
N=500	1.728974	0.3915924	0.8370398
N=1000	1.792014	0.4899758	0.8458286
N=5000	2.306566	0.5470676	0.8705451
N=10000	0.1155536	0.8836072	0.7911635
N=20000	0.9976591	0.7542213	0.8239918
N=50000	0.03140528	0.8997948	0.7887981