Limit theorems for forward and backward processes of numbers of non-empty urns in infinite urn schemes

Artyom Kovalevskii (joint work with
M. Chebunin and N. Zakrevskaya)

Novosibirsk State Technical University
Novosibirsk State University

pandorra@ngs.ru

Novosibirsk, 2022

Introduction

The motivation for this work was the procedure for writing essays by students in the Internet age: a student's essay sometimes is simply a combination of two or more texts found using a search engine. As a result, we cannot determine the student's intellectual contribution. Therefore, we need an algorithm that allows us to quickly identify the presence of heterogeneous fragments in a text. Our models and methods are completely probabilistic.

Forward and backward processes of numbers of different words

Hamlet: To be or not to be

hamlet to be or not to be

k	0	1	2	3	4	5	6	7
R_k	0	1	2	3	4	5	5	5

be to not or be to hamlet

k	0	1	2	3	4	5	6	7
R'_k	0	1	2	3	4	4	4	5

An infinite urn scheme

There is a countably infinite dictionary where the words are numbered $1, 2, \ldots$. Words are chosen one-by-one independently of each other.
Let X_i be the number of the word at ith position, $1 \le i \le n$.

$$\mathrm{P}(X_i = j) = p_j > 0, \quad j \ge 1$$

$$p_1 + p_2 + \ldots = 1$$

$$p_1 \ge p_2 \ge \ldots$$

General theorems

Denote by R_n the number of different words in the text of length n

$$\mathsf{E}R_n = \sum_{i=1}^{\infty}(1 - (1 - p_i)^n)$$

$$\mathsf{Var}R_n \le \mathsf{E}R_n$$

$$\mathsf{E}R_n \to \infty, \quad \mathsf{E}R_n/n \to 0$$

[Bahadur, 1960]

$$R_n/\mathsf{E}R_n \stackrel{a.s.}{\to} 1$$

[Karlin, 1967]

Example: Shakespeare's sonnets

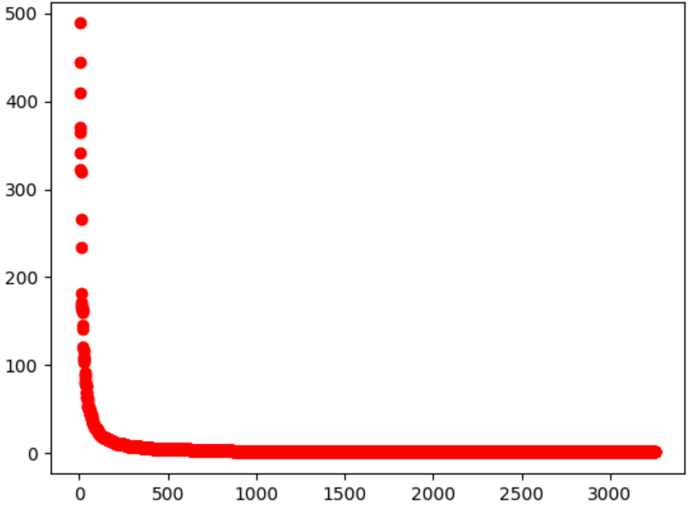
$n = 17516$, $R_n = 3258$

Most frequent words:

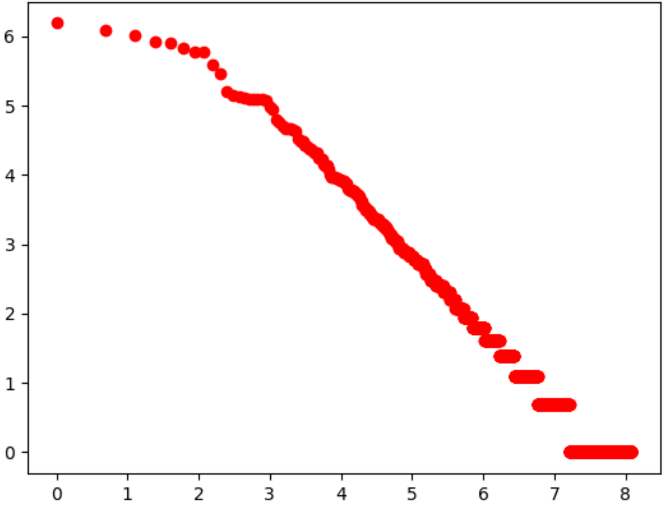
'and': 489,	'the': 444,
'to': 409,	'of': 371,
'my': 364,	'i': 341,
'in': 322,	'that': 320,
'thy': 266,	'thou': 234,
'with': 181,	'for': 171,
'is': 169,	'not': 167,
'but': 164,	'me': 164,
'a': 163,	'thee': 162,
'love': 160,	'so': 145,

— end of top 20 —

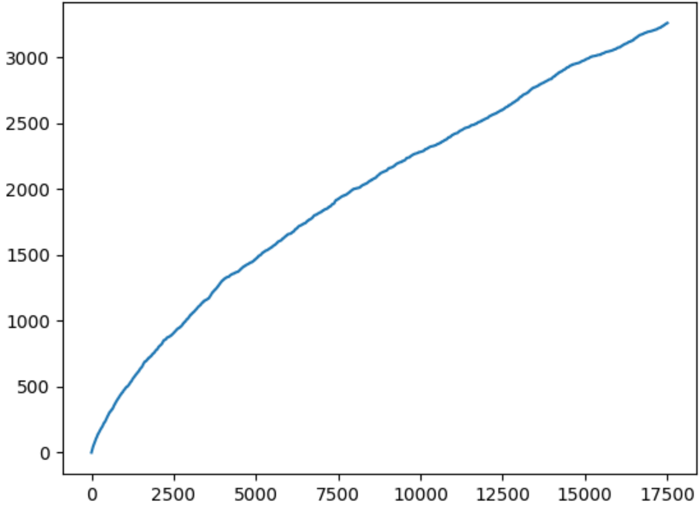
'be': 141,	'as': 121,
'all': 117,	'you': 110,
...	...



Frequences of words in Shakespeare's sonnets



Logs of frequences of words to logs of ranks in Shakespeare's sonnets (Zipfian diagram)



The process of numbers of different words in Shakespeare's sonnets (Heaps' diagram)

Remember the formula

$$\mathsf{E}R_k = \sum_{i=1}^{\infty} (1 - (1 - p_i)^k).$$

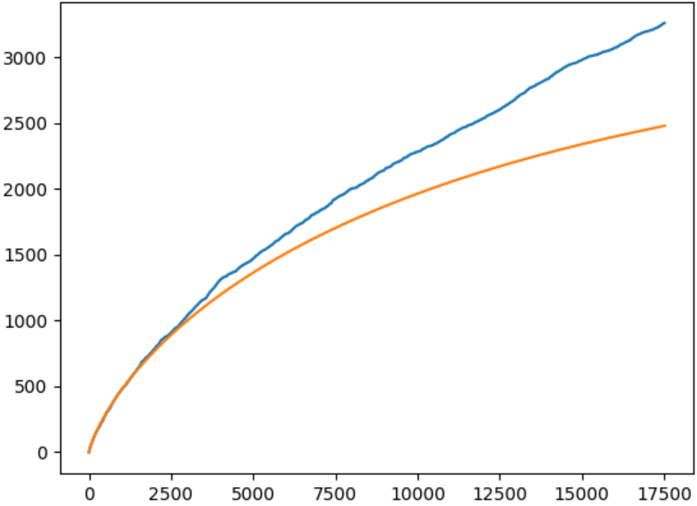
We estimate the unknown expectation by

$$R_k^* = \sum_{i=1}^{R_n} (1 - (1 - p_i^*)^k)$$

with

$$p_i^* = n_i / n,$$

n_i be the number of occurences of a word with rank i.
The next figure illustrates the badness of this approximation.



The process of R_k with its empirical approximation R_k^*

A regular case

Regularity condition:

$$\alpha(x) := \max\{k > 0 : \ p_k \geq 1/x\} = x^{\theta} L(x), \quad 0 < \theta < 1,$$

$L(\cdot)$ is the slowly varying function of the real argument:
$L(tx)/L(x) \to 1$ as $x \to +\infty$ for any real $t > 0$.

Equivalent condition:

$$p_i = i^{-1/\theta} l(i),$$

$l(\cdot)$ is the another slowly varying function.
The model is the elementary probability model that corresponds to the Zipf's Law (Zipf, 1936) of power decreasing of word probabilities.

Poissonization

Let (see Karlin (1967)) $\Pi = \{\Pi(t),\ t \geq 0\}$ be a Poisson process with parameter 1. We denote by $X_i(n)$ a number of balls in urn i. According to well-known property of splitting of Poisson flows, stochastic processes $\{X_i(\Pi(t)),\ t \geq 0\}$ are Poisson with intensities p_i and are mutually independent for different i's. The definition implies that

$$R_{\Pi(t)} = \sum_{i=1}^{\infty} \mathrm{I}(X_i(\Pi(t)) \geq 1).$$

[Theorem 1 in Karlin (1967)]

Let $\theta \in [0, 1)$. *Then* $\mathsf{E}R_{\Pi(t)} \sim \alpha(t)\Gamma(1-\theta)$ *as* $t \to \infty$.

Proof

Clearly

$$R_{\Pi(t)} = \sum_{i=1}^{\infty} \mathsf{I}(X_i(\Pi(t)) \geq 1),$$

$$\mathsf{E}R_{\Pi(t)} = \sum_{i=1}^{\infty} \mathsf{P}(X_i(\Pi(t)) \geq 1) = \sum_{i=1}^{\infty}(1 - e^{-p_i t}).$$

In view of the definition of $\alpha(x)$ we may write

$$\mathsf{E}R_{\Pi(t)} = \int_0^{\infty} (1 - e^{-t/x}) d\alpha(x).$$

Integration by parts and a change of variable yelds

$$\mathsf{E}R_{\Pi(t)} = \int_0^{\infty} \frac{t}{x^2} e^{-t/x} \alpha(x)\, dx = t \int_0^{\infty} e^{-ty} \alpha(1/y)\, dy.$$

A standard Abelian argument produces the result

$$\mathsf{E}R_{\Pi(t)} \sim \alpha(t)\Gamma(1-\theta).$$

See Theorem A6.3.1 in Borovkov (2009) with $V(t) = \alpha(1/t)$ for details.
The proof is complete.

[Theorem 1' in Karlin (1967)]
Let $\theta \in [0,1)$. *Then* $\mathsf{E}R_n \sim \alpha(n)\Gamma(1-\theta)$ *as* $n \to \infty$.

Theorems under the regularity condition

Karlin (1967): $(R_n - \mathsf{E}R_n)/\sqrt{\mathsf{Var}R_n}$ converges weakly to the standard normal distribution,

$$\mathsf{E}R_n \sim \Gamma(1-\theta)\alpha(n),$$

$$\mathsf{Var}R_n/\mathsf{E}R_n \to 2^\theta - 1,$$

$\Gamma(\cdot)$ is the Euler gamma.
So $(R_n - \mathsf{E}R_n)/\sqrt{\mathsf{E}R_n}$ converges weakly to the centered normal distribution with variance $2^\theta - 1$.
Chebunin and Kovalevskii (2016):

$$Z_n = \{Z_n(t),\ 0 \le t \le 1\} = \{(R_{[nt]} - \mathsf{E}R_{[nt]})/\sqrt{\mathsf{E}R_n},\ 0 \le t \le 1\}$$

converges weakly in $D(0,1)$ with uniform metrics to a centered Gaussian process Z_θ with continuous a.s. sample paths and covariance function

$$K(s,t) = (s+t)^\theta - \max(s^\theta, t^\theta).$$

New theorem under the regularity condition

Theorem (for joint distribution)

If the regularity condition holds then $(Z_n, Z'_n) = \{(Z_n(t), Z'_n(t)),\ 0 \leq t \leq 1\}$ *converges weakly in the uniform metrics in* $D(0,1)^2$ *to 2-dimensional Gaussian process* (Z, Z') *with zero expectation and covariance function*

$$\mathrm{E}Z(s)Z(t) = \mathrm{E}Z'(s)Z'(t) = K(s,t), \quad \mathrm{E}Z(s)Z'(t) = K'(s,t),$$

$$K(s,t) = (s+t)^\theta - \max(s^\theta, t^\theta),$$

$$K'(s,t) = ((s+t)^\theta - 1)1(s+t > 1).$$

From the Theorem we have that the limiting process $\{(Z(t) - Z'(t))/\sqrt{2},\ 0 \leq t \leq 1/2\}$ is the stochastically self-similar process which coinside in distribution with the limiting process of Durieu and Wang (2016). So the Theorem gives an alternative way to simulate these processes without additional randomization.

Corollary under the regularity condition

Corollary (for the difference of processes) *If the regularity condition holds then*

$$J_n = \frac{\sum_{k=1}^{n}(R_k - R'_k)}{n\sqrt{R_n}}$$

converges weakly to a centered normal random variable with variance $\frac{\theta}{\theta+2}$.

The Corollary gives the opportunity to test the homogeneity of the sample using any consistent estimate θ^* of parameter θ. Various classes of such estimates have been obtained and analysed by Hill (1975), Nicholls (1978), Zakrevskaya and Kovalevskii (2001, 2019), Guillou and Hall (2002), Ohannessian and Dahleh (2012), Chebunin (2014), Chebunin and Kovalevskii (2019a, 2019b), Chakrabarty et al. (2020).

The p-value is calculated using the tail of the standard normal distribution and the observed value J_{obs} of J_n:

$$\text{p-value} = 2\overline{\Phi}\left(|J_{obs}|\sqrt{1+2/\theta^*}\right).$$

Parameter's estimation

$$\theta_n = \int_0^1 \log^+ R_{[nt]}\, dA(t), \quad \theta'_n = \int_0^1 \log^+ R'_{[nt]}\, dA(t),$$

here $\log^+ x = \max(\log x,\, 0)$. Function $A(\cdot)$ has bounded variation and

$$A(0) = A(1) = 0, \quad \lim_{x \downarrow 0} \log x \int_0^x |dA(t)| = 0, \quad \int_0^1 \log t\, dA(t) = 1.$$

Let

$$\widehat{\theta} = (\theta_n + \theta'_n)/2.$$

Theorem (consistence)
Let $p_i = i^{-1/\theta} l(i, \theta)$, $\theta \in [0, 1]$, and $l(x, \theta)$ is a slowly varying function as $x \to \infty$. Then the estimator $\widehat{\theta}$ is strongly consistent.

Corollary

Let

$$A(t) = \begin{cases} 0, & 0 \le t \le 1/2; \\ -(\log 2)^{-1}, & 1/2 < t < 1; \\ 0, & t = 1. \end{cases}$$

Then

$$\theta_n = \log_2 \left(R_n / R_{[n/2]} \right),$$

$$\theta_n' = \log_2 \left(R_n / R_{[n/2]}' \right),$$

$$\widehat{\theta} = \log_2 \left(R_n / \sqrt{R_{[n/2]} R_{[n/2]}'} \right), \quad n \ge 2.$$

Zipf-Mandelbrot law

[Zipf, 1936], [Mandelbrot, 1965]

$$p_i = c(i+q)^{-1/\theta}, \quad i \geq 1, \quad 0 < \theta < 1, \quad q > -1.$$

Here

$$c = \left(\zeta(1/\theta, q+1)\right)^{-1},$$

$$\zeta(\alpha, x) = \sum_{i=0}^{\infty} (i+x)^{-\alpha}$$

is the Hurvitz zeta function.

Let

$$r(n) = \sum_{i=1}^{\infty} (1 - (1 - \widehat{p}_i)^n)$$

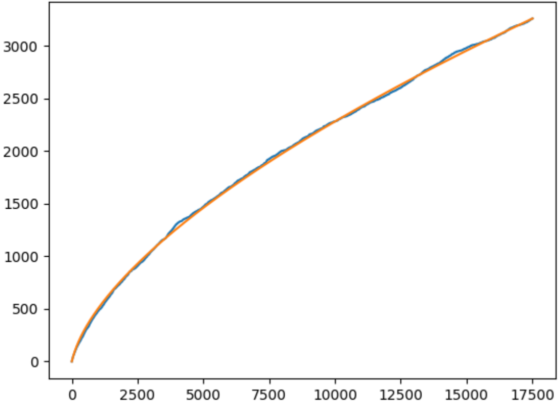
with

$$\widehat{p}_i = \widehat{c}(i + q_n)^{-1/\theta_n}, \quad i \geq 1,$$

q_n is such that $r(n) = R_n$.

Theorem *If the Zipf—Mandelbrot law is true then there is q_n such that $r(n) = R_n$ a.s., and $q_n \to q$ in probability.*

Shakespeare's sonnets



The forward process of numbers of different words for Shakespeare's sonnets and its approximation.
$n = 17516$, $R_n = 3258$, $\theta_n = 0.6267$, $q_n = 46.39$.

Omega squared statistics

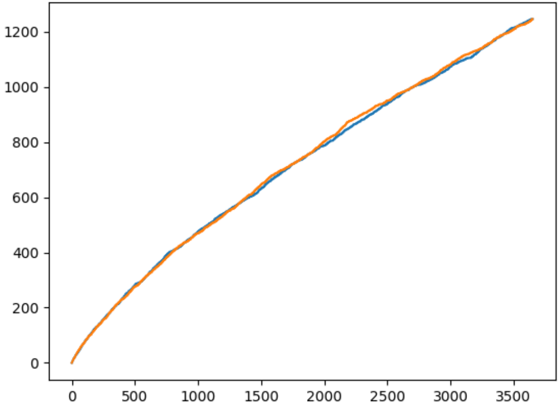
Corollary *If the regularity holds,* $0 < \theta < 1$, *then the statistics*

$$\omega_n^2 = \int_0^1 (Z_n(t) - Z_n'(t))^2 dt$$

converges weakly to a random variable ω_θ^2.

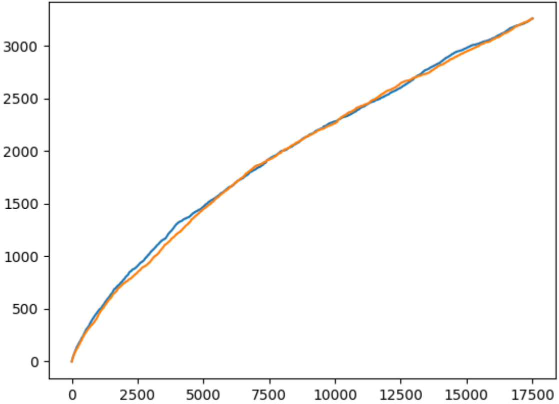
Sonnets for analysis

Thomas Wyatt (32 sonnets), 1542
William Shakespeare (154 sonnets), 1609
Charlotte Smith, ELEGIAC SONNETS (sonnets I - LIX), 1784



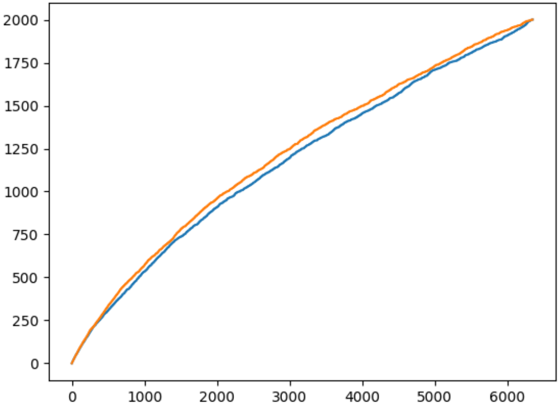
Forward and backward processes of numbers of different words for Wyatt's sonnets

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Wyatt	-0.1139	0.7556	0.7459	0.7507	0.8275	0.0681



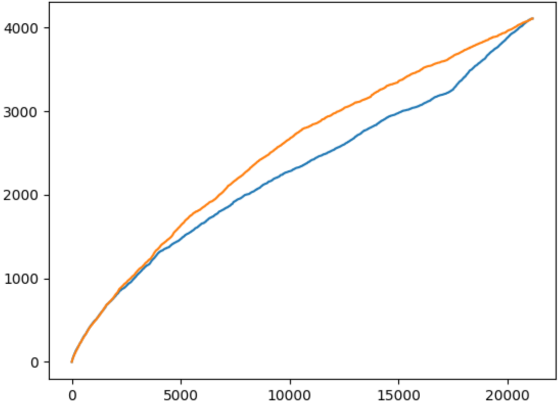
Forward and backward processes of numbers of different words for Shakespeare's sonnets

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Shakespeare	0.2939	0.6267	0.6274	0.6271	0.5475	0.3868



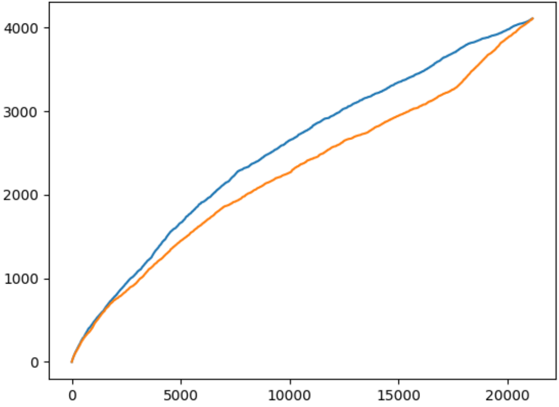
Forward and backward processes of numbers of different words for Smith's sonnets

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Smith	-0.8748	0.6788	0.62	0.6494	0.0772	0.883



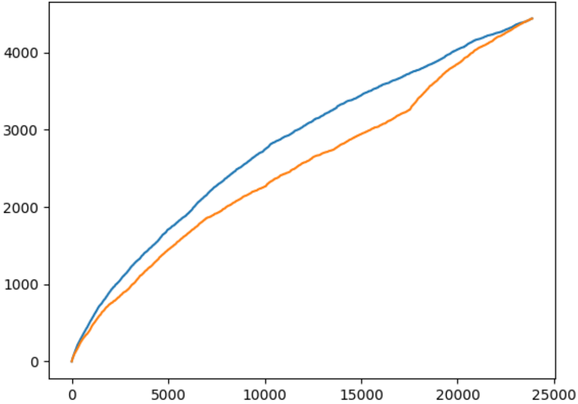
Forward and backward processes of numbers of different words for Shakespeare+Wyatt

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Shakespeare+Wyatt	-3.7886	0.8082	0.5634	0.6858	0.0000	20.3048



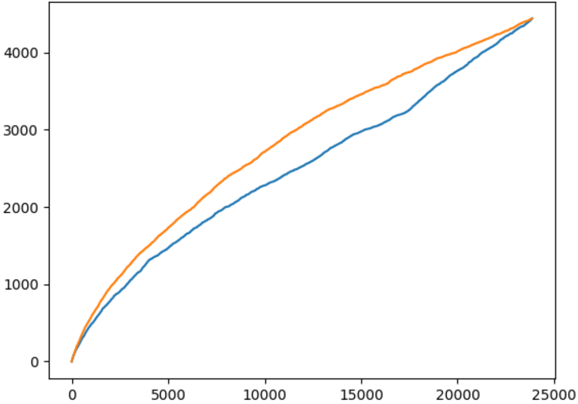
Forward and backward processes of numbers of different words for Wyatt+Shakespeare

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Wyatt+Shakespeare	4.2126	0.5837	0.7948	0.6893	0.0000	22.4295



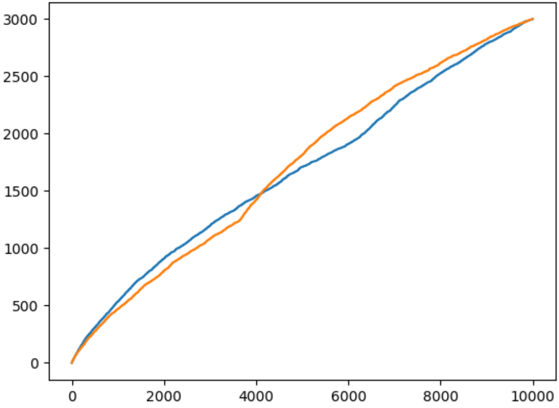
Forward and backward processes of numbers of different words for Smith+Shakespeare

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Smith+Shakespeare	4.6056	0.552	0.7925	0.6723	0.0000	27.3113



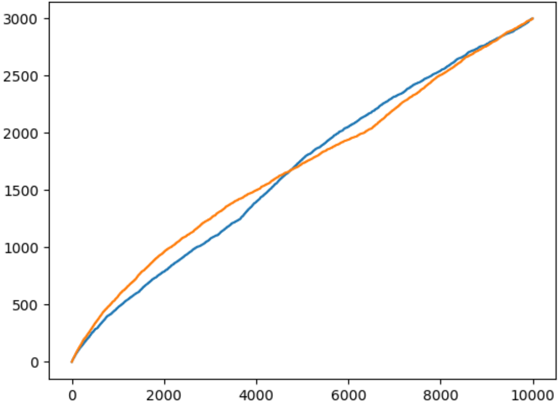
Forward and backward processes of numbers of different words for Shakespeare+Smith

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Shakespeare+Smith	-4.8183	0.8146	0.5444	0.6795	0.0000	28.7613



Forward and backward processes of numbers of different words for Smith+Wyatt

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Smith+Wyatt	-0.5909	0.8108	0.7256	0.7682	0.2620	4.5616



Forward and backward processes of numbers of different words for Wyatt+Smith

Author	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Wyatt+Smith	-0.4583	0.7627	0.7924	0.7775	0.3863	3.7753

Author(s)	J_n	θ_n	θ'_n	$\widehat{\theta}$	p-value	ω_n^2
Wyatt	-0.1139	0.7556	0.7459	0.7507	0.8275	0.0681
Shakespeare	0.2939	0.6267	0.6274	0.6271	0.5475	0.3868
Smith	-0.8748	0.6788	0.62	0.6494	0.0772	0.883
Shakespeare+Wyatt	-3.7886	0.8082	0.5634	0.6858	0.0000	20.3048
Wyatt+Shakespeare	4.2126	0.5837	0.7948	0.6893	0.0000	22.4295
Smith+Shakespeare	4.6056	0.552	0.7925	0.6723	0.0000	27.3113
Shakespeare+Smith	-4.8183	0.8146	0.5444	0.6795	0.0000	28.7613
Smith+Wyatt	-0.5909	0.8108	0.7256	0.7682	0.2620	4.5616
Wyatt+Smith	-0.4583	0.7627	0.7924	0.7775	0.3863	3.7753