Статистические свойства некоторых процедур сжатия данных

Бзикадзе Андрей Важевич, гр. 15.M03-мм

Санкт-Петербургский государственный университет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н. Некруткин В.В.
Рецензент: исследователь Советкин Е.А.

Санкт-Петербург
06 июня 2017

«Book Stack»-преобразование

Одна из рассматриваемых в ВКР процедур сжатия данных.

- $\mathbb{S} \stackrel{\text{def}}{=} \{1, 2, \ldots, S\}$ — множество книг.
- Стопка книг.
- Начальный порядок Ξ_0.
- Из стопки случайная книга перекладывается наверх.

«Book Stack»-преобразование

Одна из рассматриваемых в ВКР процедур сжатия данных.

- $\mathbb{S} \stackrel{\text{def}}{=} \{1, 2, \ldots, S\}$ — множество книг.
- Стопка книг.
- Начальный порядок Ξ_0.
- Из стопки случайная книга перекладывается наверх.

Итеративная процедура.

- $\{\eta_i\}_{i=1}^{\infty}$ — последовательность **названий** случайных книг.
- $\{\Xi_i\}_{i=0}^{\infty}$ — последовательность состояний **стопки** случайных книг.
- $\{\xi_i\}_{i=1}^{\infty}$ — последовательность **положений** случайных книг.

«Book Stack»-преобразование

Одна из рассматриваемых в ВКР процедур сжатия данных.

- $\mathbb{S} \stackrel{\text{def}}{=} \{1, 2, \ldots, S\}$ — множество книг.
- Стопка книг.
- Начальный порядок Ξ_0.
- Из стопки случайная книга перекладывается наверх.

Итеративная процедура.

- $\{\eta_i\}_{i=1}^{\infty}$ — последовательность **названий** случайных книг.
- $\{\Xi_i\}_{i=0}^{\infty}$ — последовательность состояний **стопки** случайных книг.
- $\{\xi_i\}_{i=1}^{\infty}$ — последовательность **положений** случайных книг.

$$\eta_i = \Xi_{i-1}[\xi_i].$$

Пример. $S = 5$

3
4
1
5
2

Ξ_0

Пример. $S = 5$

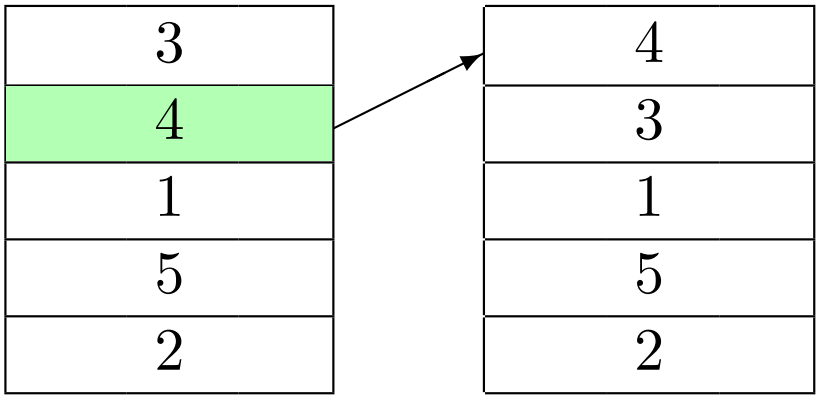
$$\eta_1 = 4$$
$$\xi_1 = 2$$

3
4
1
5
2

Ξ_0

Пример. $S = 5$

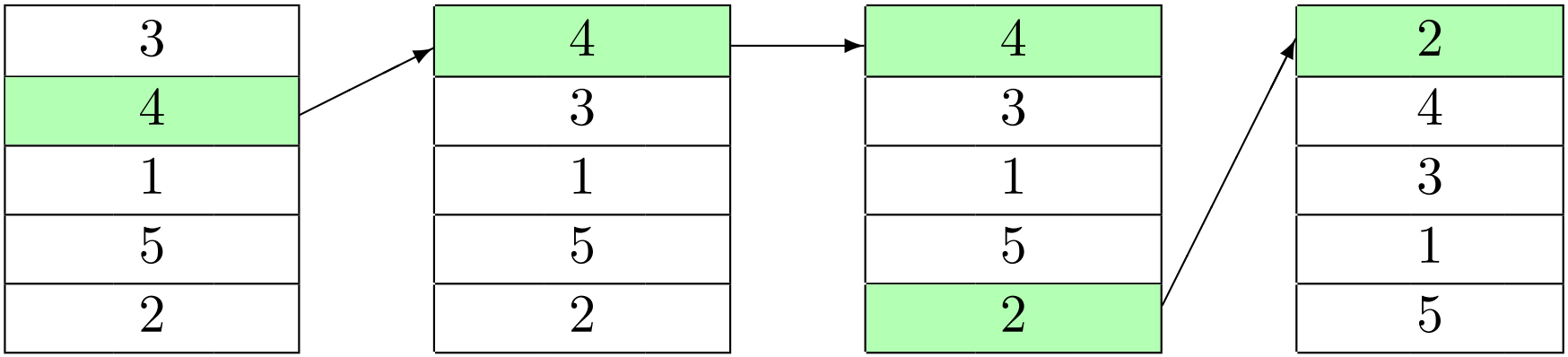
$$\eta_1 = 4$$
$$\xi_1 = 2$$



Ξ_0 Ξ_1

Пример. $S = 5$

$$\eta_1 = 4 \qquad \eta_2 = 4 \qquad \eta_3 = 2$$
$$\xi_1 = 2 \qquad \xi_2 = 1 \qquad \xi_3 = 5$$



Обзор литературы

Применение «Book Stack»-преобразования:

- Рябко Б.Я. (1980): алгоритм **сжатия** данных под названием «метод стопки книг».
- Bentley J.L. et al. (1986): тот же алгоритм под названием **«Move To Front»**.
- Рябко Б.Я. et al. (2003–2004): тест для проверки свойств **генераторов псевдослучайных чисел** под названием «Book Stack».

«Book Stack»-тест

Равносильны:

- $\mathbb{H}_0$: η_i независимы и равномерно распределены на $\mathbb{S}$.
- $\mathbb{H}_0^*$: ξ_i независимы и равномерно распределены на $\mathbb{S}$.

Два критерия для проверки $\mathbb{H}_0$: при применении к «исходной» и к «преобразованной» выборке.

Выбор альтернативной гипотезы

Альтернатива в (Бзикадзе А.В., Некруткин В.В, 2016) и бакалаврской ВКР:
$\mathbb{H}_1 : \{\eta_i\}_{i\geqslant 1}$ — н.о.р. с **неравномерным** распределением.

При больших n:

- Критерий χ^2 к $\{\eta_i\}_{i=1}^n$ (как правило) мощнее, чем к $\{\xi_i\}_{i\geqslant 1}$.
- Критерий отношения правдоподобия к $\{\eta_i\}_{i=1}^n$ мощнее, чем к $\{\xi_i\}_{i\geqslant 1}$.

Выбор альтернативной гипотезы

Альтернатива в (Бзикадзе А.В., Некруткин В.В, 2016) и бакалаврской ВКР:
$\mathbb{H}_1 : \{\eta_i\}_{i\geqslant 1}$ — н.о.р. с **неравномерным** распределением.

При больших n:

- Критерий χ^2 к $\{\eta_i\}_{i=1}^n$ (как правило) мощнее, чем к $\{\xi_i\}_{i\geqslant 1}$.
- Критерий отношения правдоподобия к $\{\eta_i\}_{i=1}^n$ мощнее, чем к $\{\xi_i\}_{i\geqslant 1}$.

В магистерской ВКР:
$\mathbb{H}_1 : \{\eta_i\}_{i\geqslant 1}$ — **эргодическая** однородная марковская цепь (ЭОМЦ) со стационарным **равномерным** на $\mathbb{S}$ распределением и переходной матрицей $\mathbf{P}^{(\eta)}$: $\operatorname{tr}\left(\mathbf{P}^{(\eta)}\right) \neq 1$.

Марковское свойство Ξ_i

Пусть входная последовательность $\{\eta_i\}_{i\geqslant 1}$ — ОМЦ с матрицей переходов $\mathbf{P}^{(\eta)} = (p_{ij})$.

Предложение

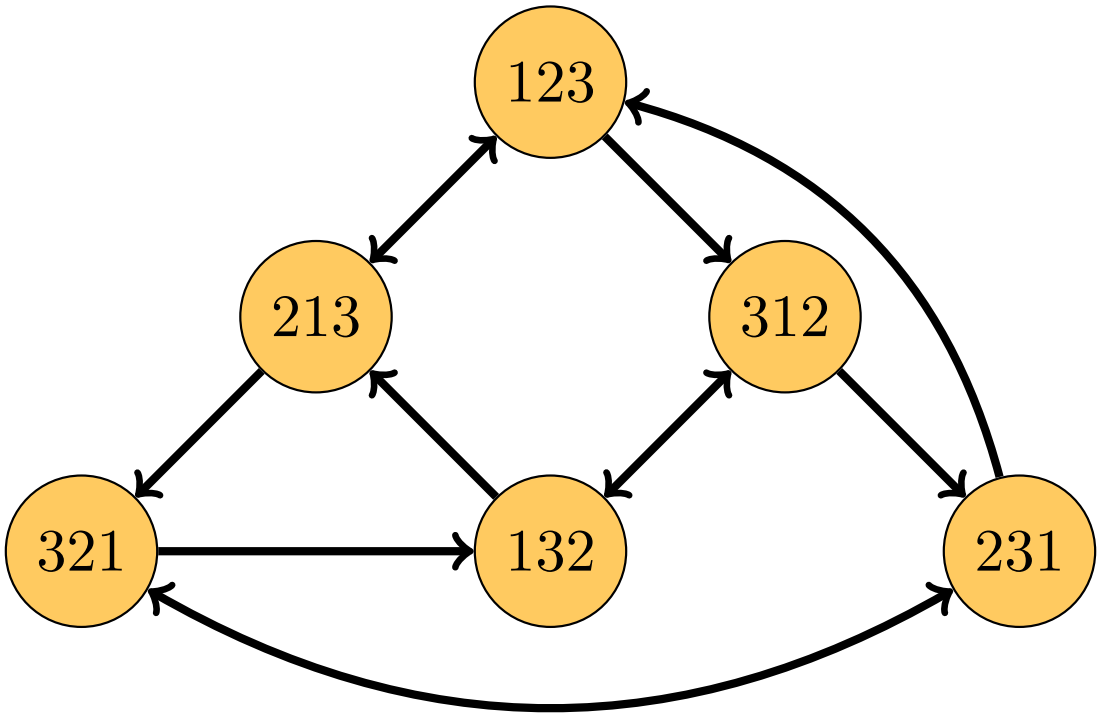
1. *Последовательность* $\{\Xi_n\}_{n\geqslant 1}$ *образует ОМЦ.*

2. *Если входная ОМЦ* $\{\eta_n\}_{n\geqslant 1}$ *—* ***эргодическая****, то последовательность* $\{\Xi_n\}_{n\geqslant 1}$ *имеет ровно* ***один непериодический эргодический класс*** *и, быть может, несколько несущественных состояний.*

3. *Если же* $p_{ij} > 0$ *при всех* i, j*, то* ***несущественных состояний нет****, т.е.* $\{\Xi_n\}_{n\geqslant 1}$ *эргодическая.*

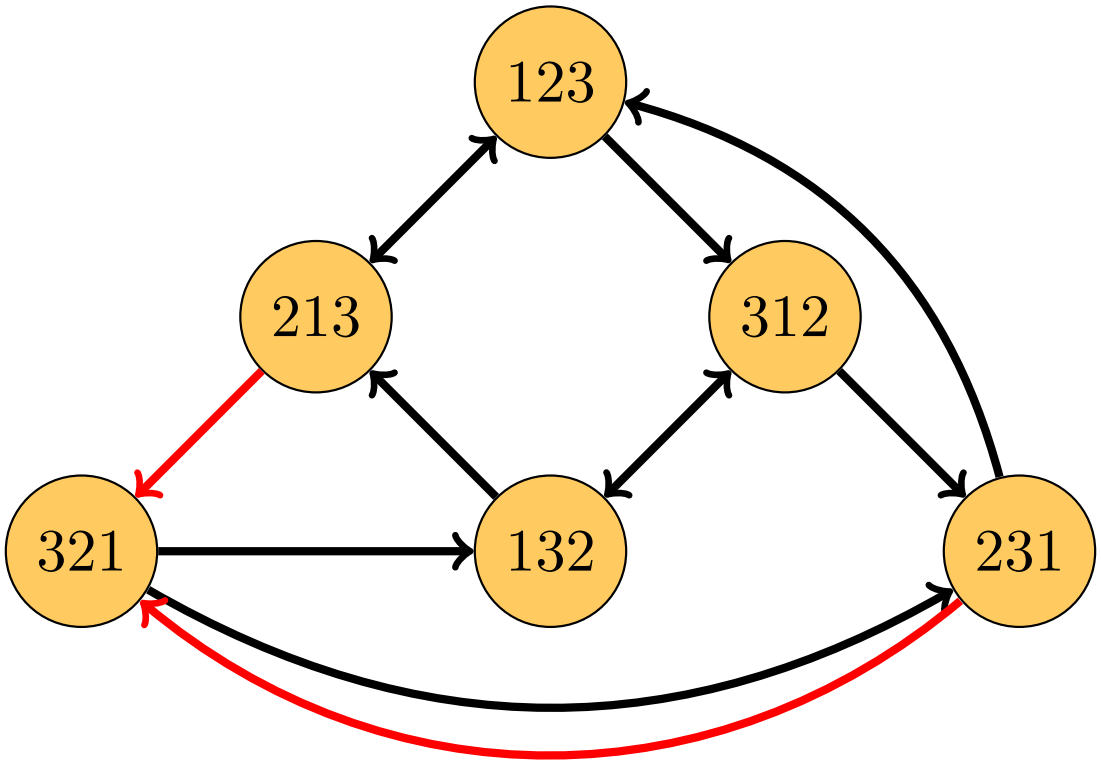
Марковское свойство Ξ_i. Пример

Рассмотрим графическое изображение ОМЦ $\{\Xi_i\}_{i\geqslant 1}$.
Входная ОМЦ $\{\eta_i\}_{i\geqslant 1}$ — эргодическая с матрицей $\mathbf{P}^{(\eta)} = (p_{ij})$.
$S = 3$, $p_{ij} > 0$ для всех i, j.



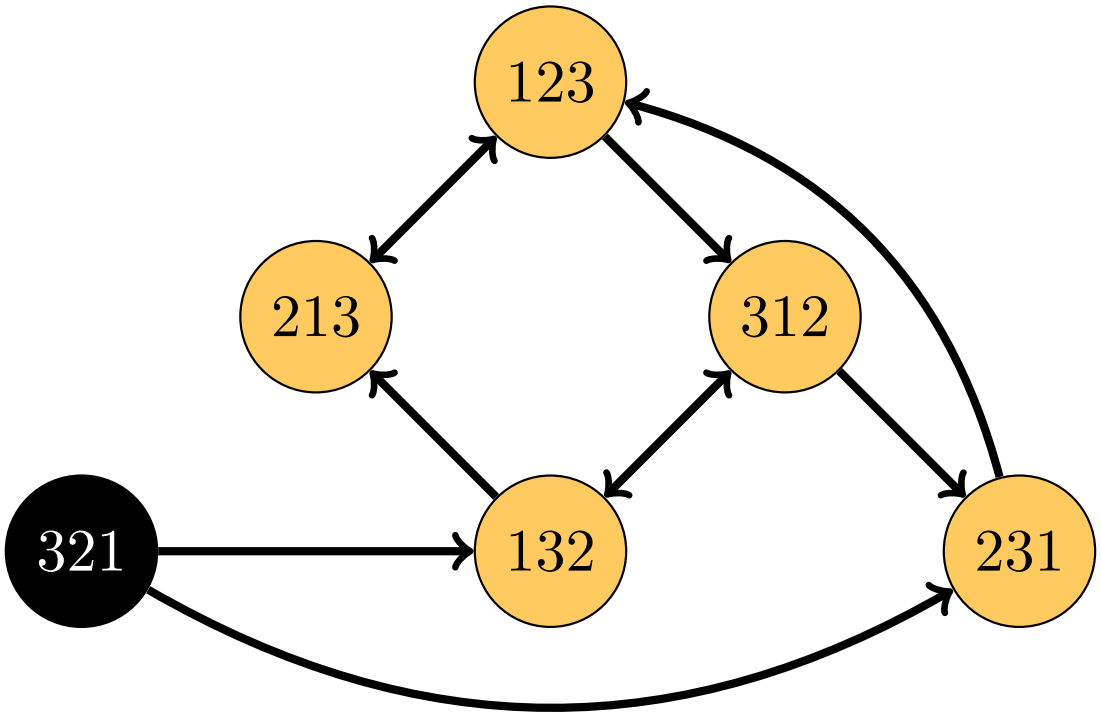
Марковское свойство Ξ_i. Пример

Рассмотрим графическое изображение ОМЦ $\{\Xi_i\}_{i\geqslant 1}$.
Входная ОМЦ $\{\eta_i\}_{i\geqslant 1}$ — эргодическая с матрицей $\mathbf{P}^{(\eta)} = (p_{ij})$.
$S = 3$, $p_{23} = 0$, $p_{33} = 0$ и остальные $p_{ij} > 0$.



Марковское свойство Ξ_i. Пример

Рассмотрим графическое изображение ОМЦ $\{\Xi_i\}_{i\geqslant 1}$.
Входная ОМЦ $\{\eta_i\}_{i\geqslant 1}$ — эргодическая с матрицей $\mathbf{P}^{(\eta)} = (p_{ij})$.
$S = 3$, $p_{23} = 0$, $p_{33} = 0$ и остальные $p_{ij} > 0$.
$(3, 2, 1)^{\mathrm{T}}$ — несущественное состояние ОМЦ $\{\Xi_i\}_{i\geqslant 1}$.



Предельное поведение ξ_i. Закон Больших Чисел

Пусть $\{\eta_i\}_{i\geqslant 1}$ — ЭОМЦ с м.п. $\mathbf{P}^{(\eta)} = (p_{ij})$.
Обозначим $(\pi_\alpha, \alpha \in \mathfrak{S}_S)$ — стационарное распределение Ξ_i,
$\tau_k = \tau_k(n) = \mathbb{I}_k(\xi_1) + \ldots + \mathbb{I}_k(\xi_n)$ для $1 \leqslant k \leqslant S$ и положим

$$s_k \stackrel{\text{def}}{=} \sum_{j=1}^{S} \sum_{\substack{\alpha\in\mathfrak{S}_S \\ \alpha_k=j}} \pi_\alpha p_{\alpha[1]j}.$$

Теорема

Для всех $k \in 1:S$ при $n \to \infty$

- $\mathbb{P}(\xi_n = k) \to s_k$.
- $\mathbb{E}\left(\tau_k/n - s_k\right)^2 = \mathrm{O}(1/n)$.

Распределение с вероятностями s_j будем обозначать $\mathcal{R}$.

Предельное поведение ξ_i. ЦПТ

Обозначим $\mathbf{s} = (s_1, \ldots, s_S)^{\mathrm{T}}$ и

$$a_{k,\ell} = \begin{cases} s_k(1 - s_k), & \text{при } k = \ell, \\ -s_k s_\ell, & \text{при } k \neq \ell. \end{cases}$$

Теорема

$$\mathcal{L}\left(\sqrt{n}\left(\tau_n^{(\xi)}/n - \mathbf{s}\right)\right) \Rightarrow \mathcal{N}\left(0, \Sigma_\xi\right),$$

где матрица Σ_ξ размеров $S \times S$ имеет компоненты

$$\left(\Sigma_\xi\right)_{k,\ell} = a_{k,\ell} + C_{k,\ell},$$

а $C_{k,\ell}$ некоторые константы, зависящие только от $\mathbf{P}^{(\eta)}$.

Сравнение предельных распределений η_i и ξ_i

Теорема

Пусть $\{\eta_i\}_{i\geqslant 1}$ образуют ЭОМЦ со стационарным ***равномерным*** *распределением и матрицей переходных вероятностей $\mathbf{P}^{(\eta)}$: $\operatorname{tr}\left(\mathbf{P}^{(\eta)}\right) \neq 1$. Тогда предельное распределение $\mathcal{R}$ не совпадает с равномерным.*

Без предположения, что $\operatorname{tr}\left(\mathbf{P}^{(\eta)}\right) \neq 1$, утверждение Теоремы, вообще говоря, неверно.

Сравнение критериев χ^2 при применении к η_i и ξ_i

Альтернатива:
$\mathbb{H}_1$: $\{\eta_i\}_{i\geqslant 1}$ — **эргодическая** однородная марковская цепь со стационарным **равномерным** распределением и переходной матрицей $\mathbf{P}^{(\eta)}$: $\operatorname{tr}\left(\mathbf{P}^{(\eta)}\right) \neq 1$.

Теорема

Критерий χ^2 против альтернативы $\mathbb{H}_1$

- *При применении к «входным» $\{\eta_i\}_{i\geqslant 1}$ — несостоятельный.*
- *При применении к «выходным» $\{\xi_i\}_{i\geqslant 1}$ — состоятельный.*

Пример. Модель

Модель марковской цепи:

- Задано: $0 < \delta < 1$.
- Матрица переходных вероятностей: $\mathbf{P} = (p_{ij})$, где $p_{ii} = \delta$ и $p_{ij} = (1 - \delta)/(S - 1)$ при $i \neq j$.

Моделирование:

- Вихрь Мерсенна.
- n — размер выборки, m — количество выборок.
- Критерий χ^2 с $S - 1$ степенью свободы: m штук P-значений.

Цель:

- Сравнение мощности критерия χ^2 «до» и «после» преобразования «Book Stack».

Пример. Моделирование

Параметры: $S = 3$, $n = 10^4$, $\delta = 1/S + 0.01$, $m = 100$.

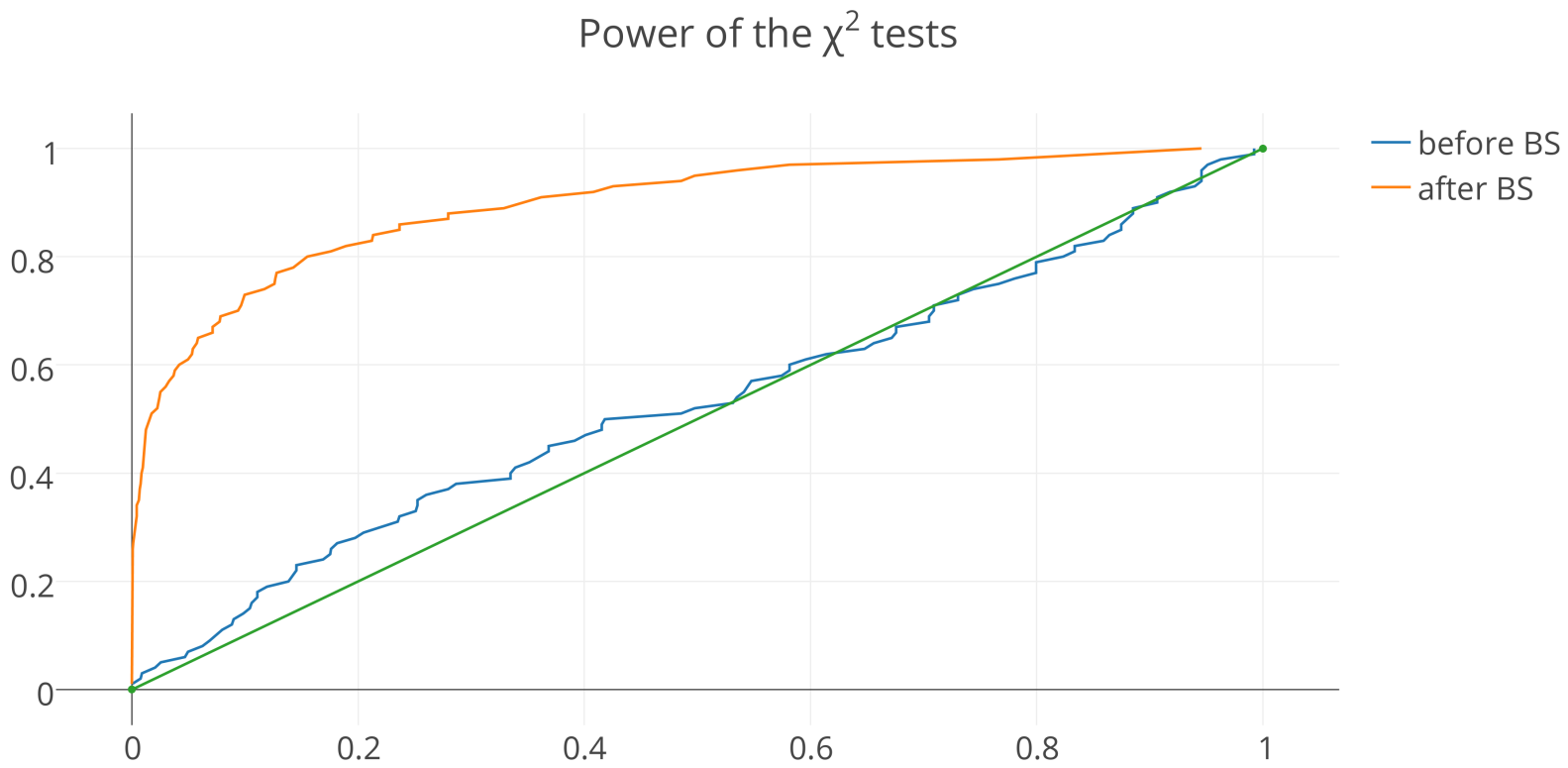


Рис.: Мощности критериев χ^2 до/после «Book Stack». Для χ^2 «до» P-значение критерия Колмогорова-Смирнова равно 0.48, «после» — меньше $2.2 \cdot 10^{-16}$.

Итоги

- Получено обобщение теоретико вероятностных результатов бакалаврской ВКР на случай, когда $\{\eta_i\}_{i\geqslant 1}$ образуют ЭОМЦ.
- При альтернативной гипотезе, что $\{\eta_i\}_{i\geqslant 1}$ — ЭОМЦ с матрицей переходов $\mathbf{P}^{(\eta)}$: $\operatorname{tr}\left(\mathbf{P}^{(\eta)}\right) \neq 1$, критерий χ^2 «до» преобразования несостоятельный, а такой же критерий «после» — состоятельный.

Также получены следующие результаты

- Рассмотрено **обобщение** «Book Stack»-преобразования, при котором **сохраняются** многие теоретико вероятностые результаты.
- Рассмотрен другой тест «Order» для проверки той же гипотезы $\mathbb{H}_0$ и теоретически обоснована **бесперспективность** его применения.