Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»
при Вычислительном центре им. А. А. Дородницына РАН

Целых Влада Руслановна

# Статистические критерии адекватности вероятностных тематических моделей коллекции текстовых документов

010900 — Прикладные математика и физика

Бакалаврская диссертация

**Научный руководитель:**
ст.н.с. ВЦ РАН, д.ф.-м.н.
Воронцов Константин Вячеславович

Москва
2013 г.

# Содержание

**Аннотация**

Работа посвящена построению критерия, проверяющего одно из основных предположений тематического моделирования — гипотезу условной независимости слов в теме от документа. Предлагаются два статистических теста: один основан на вычислении эмпирических распределений статистики хи-квадрат путём сэмплирования, а второй — на множественном использовании точного теста Фишера. Рассматривается применение предложенных тестов для проверки адекватности вероятностных тематических моделей.

# 1 Введение

Тематическое моделирование (topic modeling) — одно из активно развивающихся приложений машинного обучения к анализу текстов [9]. *Тематическая модель* коллекции текстовых документов определяет, к каким темам относится каждый документ, и какие термины образуют каждую тему. *Вероятностная тематическая модель* описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем. Это позволяет решать задачи классификации, кластеризации и категоризации текстов, а также создавать тематические поисковые системы, позволяющие по тексту произвольной длины находить документы схожей тематики. Кроме того, тематические модели широко используются в области компьютерного зрения: для классификации изображений [1], определения подписей к ним [2] и построения иерархий [3, 4].

Исходными данными для тематической модели является множество (коллекция) текстовых документов $D$ и множество (словарь) терминов $W$. Каждый документ $d \in D$ представляется последовательностью терминов $(w_1, \ldots, w_{n_d})$ из $W$, где $n_d$ — длина документа. Через $n_{dw}$ обозначается число вхождений термина $w$ в документ $d$.

*Латентно-семантический анализ* [5] (латентно-семантическое индексирование в информационном поиске)—метод обработки информации, анализирующий взаимосвязь между документами и встречающимися в них терминами путем представления документов и терминов в пространстве так называемых “тем”. Для этого используется сингулярное разложение матрицы слов-на-документы $A$, обычно содержащей в качестве элементов веса, учитывающие частоты использования каждого термина в каждом документе и участие термина во всех документах (tf-idf). Иначе говоря, матрица $A$ представляется в виде произведения трех матриц:

$$A = U\Lambda V^T,$$

где $U$ и $V$—ортогональные матрицы, а $\Lambda$—диагональная матрица, на диагонали которой—собственные значения $AA^T$, называемые сингулярными значениями матрицы $A$. Оставляя $k$ наибольших диагональных элементов матрицы $\Lambda$ и взяв соответствующие им столбцы матриц $U$ и $V$, получается матрица

$$A_k = U_k\Lambda_k V_k^T$$

ранга $k$, которая является лучшей аппроксимацией исходной матрицы $A$ среди матриц заданного ранга $k$ (минимизирует норму Фробениуса разности матриц [16]). Та-

ким образом, каждый термин и документ представляется в общем пространстве размерности $k$, что позволяет определять близость между документами, терминами или документами и терминами как косинус угла между соответствующими векторами. Латентно-семантический анализ используется для классификации документов, кластеризации, поиска информации, позволяет решать проблемы, связанные с синонимией терминов [6].

*Вероятностный латентно-семантический анализ*, появившийся в 1999 году [10], является дальнейшим развитием *латентно-семантического анализа*, в отличие от которого имеет строгое статистическое обоснование и определяет модель порождения коллекции документов. Кроме того, данный метод позволяет разделять различные значения слов, тем самым решая проблемы, связанные с полисимией (многозначностью). Вероятностный латентный семантический анализ основывается на представлении вероятности появления пары "документ-термин" следующим образом:

$$p(d,w) = \sum_{t\in T} p(t)p(w|t)p(d|t) = \sum_{t\in T} p(d)p(w|t)p(t|d) = \sum_{t\in T} p(w)p(t|w)p(d|t),$$

где $t$ — скрытая переменная(тема). Неизвестные вероятности $p(w|t)$ и $p(t|d)$ для всех $w \in W, t \in T, d \in D$ определяются из решения задачи максимизации логарифма правдоподобия выборки:

$$\sum_{d\in D}\sum_{w\in W} n_{dw} \log \sum_{t\in T} p(w|t)p(t|d) \longrightarrow \max_{\{p(w|t)\},\{p(t|d)\}}$$

при ограничениях:

$$\begin{cases} p(w|t) \geqslant 0 & \forall w \in W, t \in T \\ p(t|d) \geqslant 0 & \forall t \in T, d \in D \\ \sum\limits_{w\in W} p(w|t) = 1 & \forall t \in T \\ \sum\limits_{t\in T} p(t|d) = 1 & \forall d \in D \end{cases}$$

Для решения данной задачи применяется итерационный процесс, каждая итерация которого состоит из двух шагов (EM-алгоритм [11]).

Вероятностная тематическая модель с априорными распределениями Дирихле была предложена Дэвидом Блэем и др. в [12] и названа *латентным размещением Дирихле* (Latent Dirichlet Allocation, LDA). Одним из ключевых предположений модели является то, что распределения тем в документах подчиняются распределению Дирихле. При этом вводится двухуровневая модель порождения каждого документа коллекции $d \in D$, описанная в Алгоритме 1.1. Каждая тема представляет собой

**Алгоритм 1.1.** Порождение документа $d$ в коллекции.

**Вход:** $p(w \mid t)$ для всех $w \in W, t \in T$, $\alpha$—параметр распределения Дирихле;

**Выход:** последовательность слов документа $d$;

1: выбрать длину документа $n_d$;
2: выбрать распределение тем в документе $p(t \mid d) \sim Dir(\alpha)$;
3: **для всех** $i = 1 \ldots n_d$
4: выбрать тему $t \sim p(t|d)$;
5: выбрать слово $w \sim p(w|t)$;

распределение вероятностей над словами $p(w|t)$, а каждый документ—распределение вероятностей над темами $p(t|d)$.

Одним из методов решения задачи тематического моделирования LDA является сэмплирование Гиббса, предложенное в [13](Gibbs Sampling, GS). В [14] описан обзор и анализ основных моделей обучения LDA, а в [15]— строгий вывод формул LDA-GS.

Все вероятностные модели основаны на следующих предположениях [10, 12].

Во-первых, предполагается, что для выявления тематики достаточно знать, какие термины встречаются в каких документах, но не важен ни порядок терминов в документах (*гипотеза «мешка слов»*), ни порядок документов в коллекции (*гипотеза «мешка документов»*). Другими словами, предполагается, что тематику документа можно узнать даже после случайной перестановки терминов, хотя для человека такой текст теряет смысл.

Во-вторых, предполагается, что существует конечное множество тем $T$ и дискретное распределение $p(d, w, t)$ на $D \times W \times T$, порождающее последовательность независимых наблюдений — троек $(d_i, w_i, t_i)$, $i = 1, \ldots, n$. Переменная $t$ является латентной (скрытой), и наблюдаемая коллекция документов представляет собой последовательность пар $(d_i, w_i)$, $i = 1, \ldots, n$, оставшихся после отбрасывания всех тем.

В-третьих, предполагается, что условное распределение вероятностей терминов $p(w \mid d, t)$ в любом документе $d$ зависит только от темы $t$, но не от самого документа. Это предположение называется *гипотезой условной независимости*:

$$p(w \mid d, t) = p(w \mid t). \tag{1.1}$$

Его можно представить в трех эквивалентных вариантах:

$$\begin{aligned} p(w \mid d,t) &= p(w \mid t), \\ p(d \mid w,t) &= p(d \mid t), \\ p(w \mid t)p(d \mid t) &= p(w,d \mid t). \end{aligned} \tag{1.2}$$

Согласно формуле полной вероятности и гипотезе условной независимости:

$$p(w \mid d) = \sum_{t \in T} p(w \mid t)p(t \mid d). \tag{1.3}$$

Построить тематическую модель коллекции — означает по известной левой части $p(w \mid d) = n_{dw}/n_d$ найти неизвестные условные распределения в правой части: $p(w \mid t)$ для каждой темы $t \in T$ и $p(t \mid d)$ для каждого документа $d \in D$, а также определить оптимальное число тем $|T|$.

Большинство тематических моделей [10, 12, 13, 14] оценивают вероятности тем $p(t \mid d, w)$ для каждого слова $w$ в каждом документе $d$. Зная эти вероятности, возможно оценить число троек:

$n_{dwt} = n_{dw}p(t \mid d, w)$ — в которых термин $w$ документа $d$ связан с темой $t$,

$n_{dt} = \sum\limits_{w \in W} n_{dwt}$ — в которых термин документа $d$ связан с темой $t$,

$n_{wt} = \sum\limits_{d \in D} n_{dwt}$ — в которых термин $w$ связан с темой $t$,

$n_t = \sum\limits_{d \in D} \sum\limits_{w \in W} n_{dwt}$ — связанных с темой $t$,

и затем по ним найти частотные оценки искомых условных вероятностей:

$$\begin{gathered} \hat{p}(t \mid d) = \frac{n_{dt}}{n_d}, \qquad \hat{p}(w \mid t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(w \mid d,t) = \frac{n_{dwt}}{n_{dt}}, \\ \hat{p}(d \mid w,t) = \frac{n_{dwt}}{n_{wt}}, \qquad \hat{p}(d \mid t) = \frac{n_{dt}}{n_t}. \end{gathered} \tag{1.4}$$

Выполнение гипотезы условной независимости (1.2) является важным требованием к вероятностной тематической модели. В [17] предлагается критерий, оценивающий степень несоответствия темы $t$ гипотезе условной независимости. Он основан на дивергенции Кульбака-Лейблера и может быть вычислен в ЕМ-алгоритме:

$$\mathrm{KL}_t = \mathrm{KL}(\hat{p}(d,w \mid t) \| \hat{p}(d \mid t)\hat{p}(w \mid t)) = \sum_{d,w} \frac{n_{dwt}}{n_t} \ln \frac{n_{dwt}n_t}{n_{dt}n_{wt}} \tag{1.5}$$

Идея состоит в генерации нескольких модельных коллекций (например, 20), для которых гипотеза условной независимости заведомо выполняется, и вычислении значений $\mathrm{KL}_t^i$ $(i = 1, 2 \ldots 20)$ по формуле (1.5) для каждой из них. Степень несоответствия

темы $t$ гипотезе условной независимости определяется на основе сравнения $\mathrm{KL}_t$ и среднего значения $\mathrm{KL}_t^i$ по модельным коллекциям.

В работе предлагаются критерии для проверки гипотезы условной независимости (1.2) в трех вариантах: для заданной пары слово–тема $(w,t)$, пары документ–тема $(d,t)$ и для заданной темы $t$. Каждый из этих вариантов имеет большую область применения в построении и оценивании вероятностных тематических моделей. Проверка гипотезы условной независимости для каждой темы $t$ дает возможность определить темы, требующие разбиения на подтемы; для каждой пары слово–тема $(w,t)$ позволяет найти слова, неоднородные в данной теме; для каждой пары документ–тема $(d,t)$—документы, которые не описываются тематической моделью. С помощью данных критериев при построении тематической модели можно сформировать начальные приближения для новых тем и решить задачу определения оптимального числа тем.

Первый рассматриваемый в работе подход к проверке гипотезы условной независимости для пары $(d,t)$ заключается в выяснении, является ли $\hat{p}(w\,|\,d,t)$ — распределение слов в документе, относящихся к теме, эмпирическим распределением выборки, порожденной из распределения слов в теме $\hat{p}(w\,|\,t)$ (оба распределения оцениваются согласно (1.4) в процессе построения тематической модели). Проверка гипотезы для пары $(w,t)$ производится аналогичным образом (отличия только в сравниваемых распределениях: $\hat{p}(d\,|\,t)$ и $\hat{p}(d\,|\,w,t)$).

Критерий хи-квадрат Пирсона — один из статистических тестов, применяемый для проверки согласия экспериментальных данных с теоретическим распределением. Он имеет свои границы применимости и, в частности, плохо подходит для разреженных распределений [8, 7], когда число возможных значений наблюдаемой переменной значительно превосходит число наблюдений, либо когда многие значения имеют крайне низкие, хотя и ненулевые, вероятности. В этих случаях распределение статистики хи-квадрат уже не описывается классической асимптотикой, и может зависеть от длины выборки и вида теоретического распределения.

Разреженные распределения естественным образом возникают в прикладных задачах статистического анализа текстов. В работе предлагается эффективный способ оценивания функции распределения и квантилей статистики хи-квадрат, основанный на сэмплировании Монте-Карло.

Рассматривается еще один подход к проверке гипотезы условной независимости, который основывается на проверке отсутствия связей между каждым словом

и документом в теме с помощью точного теста Фишера. Главным преимуществом по сравнению с предыдущим тестом является его вычислительная эффективность: проверка гипотезы с помощью теста Фишера требует значительно меньше времени, чем проверка, основанная на статистике хи-квадрат.

# 2 Критерии согласия для разреженных распределений

## 2.1 Критерий согласия хи-квадрат

Пусть имеется выборка $n$ независимых наблюдений $\{x_1, \ldots, x_n\}$ случайной величины, принимающей значения из конечного множества $\Omega$. Её эмпирическое распределение определяется как доля наблюдений $x_i$, равных $x$:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} [x_i = x], \quad x \in \Omega.$$

Критерий хи-квадрат проверяет гипотезу о том, что случайная величина имеет заданное распределение $p(x)$, $x \in \Omega$. Для этого вычисляется статистика хи-квадрат:

$$X^2 = n \sum_{x \in \Omega} \frac{(\hat{p}(x) - p(x))^2}{p(x)}. \tag{2.1}$$

Распределение статистики $X^2$ стремится к распределению хи-квадрат с $k = |\Omega| - 1$ степенями свободы: $X^2 \sim \chi^2(k)$. Нулевая гипотеза отвергается на уровне значимости $\alpha$, если значение статистики превышает $(1 - \alpha)$-квантиль этого распределения: $X^2 > \chi^2_{1-\alpha}(k)$.

Считается, что асимптотика хи-квадрат применима, если объём выборки $n \geqslant 50$ и ожидаемое число наблюдений $np(x) \geqslant 5$ для каждого $x \in \Omega$. В случаях *разреженных* распределений $p(x)$, когда вероятности $p(x)$ малы для многих $x \in \Omega$ или когда $|\Omega| \gg n$, второе условие может не выполняться даже на очень больших выборках [8]. Стандартная рекомендация — объединять значения $x \in \Omega$ в группы — для сильно разреженных распределений оказывается неприемлемой, так как результат существенно зависит от способа группирования, который выбирается произвольно.

В качестве иллюстрации рассматривается распределение, называемое *законом Ципфа*:

$$p(x) = Ax^{-s}, \quad x \in \Omega = \{1, \ldots, v\}, \tag{2.2}$$

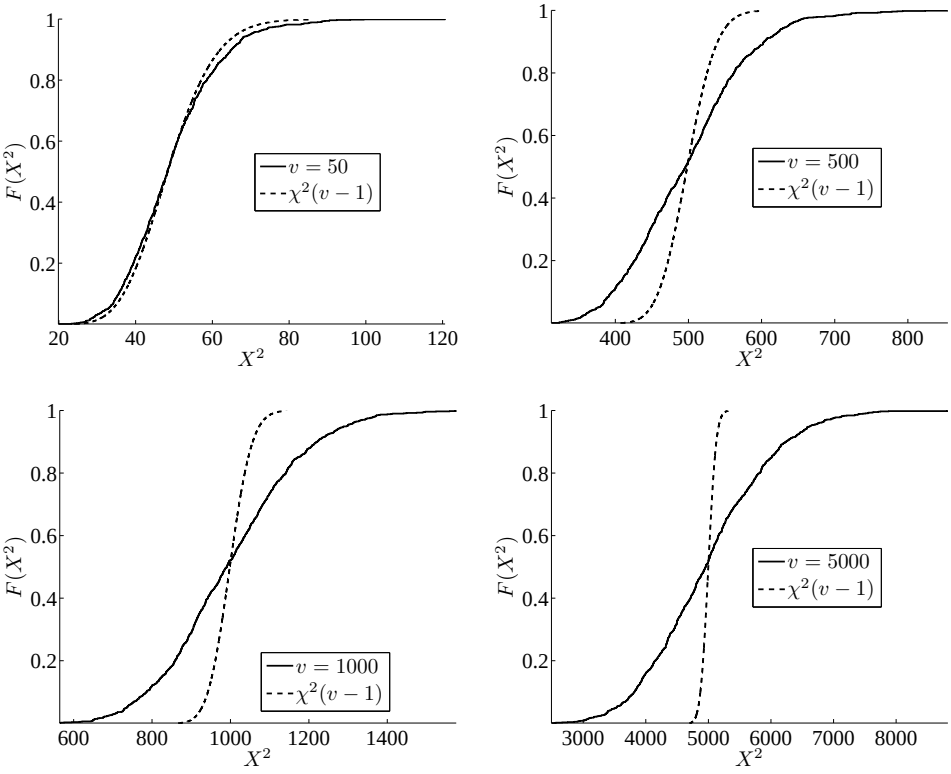


Рис. 1: Функции распределения статистики $X^2$ при $s = 1$, $v = 50,\ 500,\ 1000,\ 5000$, $n = 100$, $N = 1000$ и соответствующие функции распределения $\chi^2(v-1)$.

где $A = \left(\sum_{x=1}^{v} x^{-s}\right)^{-1}$ — нормировочный множитель, $s$ — параметр. Этот закон неплохо описывает частоты слов в текстах на естественных языках, если за $x$ принимать номера слов, упорядоченных по убыванию частоты. Параметр $s$ зависит от языка и от корпуса текстов, по которому делается оценка, но в большинстве экспериментов значение $s$ близко к 1 и находится в пределах от 0.9 до 1.2 [19, 20].

Чем больше значение параметра $s$ и размер словаря $v$, тем более разрежено распределение $p(x)$. Проводится простой вычислительный эксперимент. Рассматриваются типичные значения параметра $s = 1$ и размера словаря $v \in \{50,\ 500,\ 1000,\ 5000\}$. Генерируется $N = 1000$ выборок (искусственных текстов) длины $n = 100$ из распределения (2.2), и для каждой выборки вычисляется значение статистики $X^2$.

На рис. 1 сплошными линиями показаны эмпирические распределения статистики $X^2$, пунктирными линиями — распределения $\chi^2(v-1)$. Чем больше размер словаря, тем сильнее разрежено распределение $p(x)$, и тем сильнее отличаются $(1-\alpha)$-квантили этих распределений (при типичном значении $\alpha = 0.05$).

Таким образом, распределение хи-квадрат не может быть использовано в практических задачах анализа текстов, когда требуется проверить, является ли заданный

текст $\hat{p}(x)$ случайной выборкой из корпуса текстов $p(x)$.

## 2.2 Тест на основе сэмплирования

Для разреженных распределений $p(x)$ предлагается вместо асимптотического распределения $\chi^2(k)$ статистики $X^2$ использовать эмпирическое распределение.

**Построение теста.** Генерируется $N$ независимых выборок объёма $n$ из заданного дискретного распределения $p(x)$. Для каждой выборки вычисляется эмпирическое распределение $\hat{p}_j(x)$, $j = 1, \ldots, N$ и значение статистики $X_j^2$ по формуле (2.1). По полученным значениям $X_1^2, \ldots, X_N^2$ строится эмпирическая функция распределения статистики

$$\hat{F}_n(X^2) = \frac{1}{N} \sum_{j=1}^{N} \left[ X_j^2 < X^2 \right]$$

и вычисляется её $(1-\alpha)$-квантиль $\hat{F}_{n,1-\alpha}$. Число $N$ рекомендуется брать не менее 1000, если необходимо оценивать всю функцию распределения. Однако если оценивается только одна квантиль, $N$ можно брать порядка нескольких десятков [7].

**Применение теста.** Пусть задана выборка объёма $n$, по которой построено эмпирическое распределение $\hat{p}(x)$ и вычислено значение статистики $X^2$ согласно (2.1). Если $X^2 > \hat{F}_{n,1-\alpha}$, то нулевая гипотеза о том, что данная выборка порождена распределением $p(x)$, отклоняется.

**Рекуррентное построение теста.** Как будет показано ниже, в случае разреженных распределений значение квантили $\hat{F}_{n,1-\alpha}$ может зависеть от объёма выборки $n$. Строить тест заново для каждой выборки довольно накладно. Поэтому предлагается рекуррентный метод, позволяющий при заданном распределении $p(x)$ вычислить квантили для всех значений $n$ один раз, и затем быстро осуществлять проверку нулевой гипотезы для выборок различного объёма $n$.

В рекуррентном методе $N$ выборок $\{x_{j1}, \ldots, x_{jn}\}$ наращиваются одновременно, где $j = 1, \ldots, N$ — номер выборки, $n = 1, \ldots, n_{\max}$ — объём выборки. Для каждого $j$ строится эмпирическая гистограмма $H_j(x) = n\hat{p}_j(x)$. При добавлении каждого нового наблюдения $\xi = x_{j,n+1}$, сэмплированного из распределения $p(x)$, обновляется гистограмма и пересчитывается значение статистики $X_{j,n+1}^2$ по значению $X_{j,n}^2$. Сэмплированные выборки не сохраняются. В процессе работы алгоритм формирует двумерный массив значений статистики $X_{j,n}^2$ и одномерный массив эмпирических гистограмм $H_j(x)$. В случае $|\Omega| \gg n_{\max}$ для хранения эмпирических гистограмм лучше

**Алгоритм 2.1.** Построение теста путём рекуррентного вычисления значений статистики $X^2$ по $N$ одновременно растущим выборкам объёма $n$.

**Вход:** $p(x)$, $N$, $n_{\max}$, $\alpha$;

**Выход:** $\hat{F}_{n,1-\alpha}$ для всех $n = 1, \ldots, n_{\max}$;

**для всех** $j := 1, \ldots, N$
  сэмплировать первый элемент $j$-й выборки $\xi \sim p(x)$;
  инициализировать эмпирическую гистограмму для $j$-й выборки:
  $H_j(x) := [x = \xi]$ для всех $x \in \Omega$;
  инициализировать значение статистики $X^2$ для $j$-й выборки:
  $X^2_{j1} := 1/p(\xi) - 1$;
**для всех** $n := 1, \ldots, n_{\max} - 1$
  **для всех** $j := 1, \ldots, N$
    сэмплировать $(n+1)$-й элемент $j$-й выборки $\xi \sim p(x)$;
    обновить эмпирическую гистограмму для $j$-й выборки:
    $H_j(\xi) := H_j(\xi) + 1$;
    обновить значение статистики $X^2$ для $j$-й выборки:
    $$X^2_{j,n+1} := \frac{nX^2_{j,n} + 1}{n+1} + \frac{2H_j(\xi) - 1}{(n+1)p(\xi)} - 2;$$
**для всех** $n := 1, \ldots, n_{\max}$
  упорядочить $X^2_{1,n}, \ldots, X^2_{N,n}$ по возрастанию;
  $\hat{F}_{n,1-\alpha} := X^2_{N(1-\alpha),n}$;

использовать специальные структуры данных — разреженные векторы, не выделяющие память под нулевые значения $H_j(x)$. В таком случае расход памяти для данного алгоритма составляет $O(n_{\max}N)$; вычислительная сложность $O(n_{\max}N \log N)$. Детали реализации показаны в Алгоритме 2.1.

## 2.3 Регрессионный тест

Рассматривается частная постановка задачи: проверяется нулевая гипотеза о том, что выборка с эмпирическим распределением $\hat{p}(x)$ порождена распределением Ципфа (2.2) с параметром $s$. Распределение статистики $X^2$ строится с помощью сэмплирования и исследуется зависимость квантиля $\hat{F}_{n,1-\alpha}$ от параметров $n$, $s$ и $v$.

На рис. 2 показана зависимость 0.95-квантиля от объёма выборки $n$ и её интерполяция функцией $\tilde{F}_{1-\alpha}(n) = A + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4}$ с параметрами $A, B, C, D, E$.

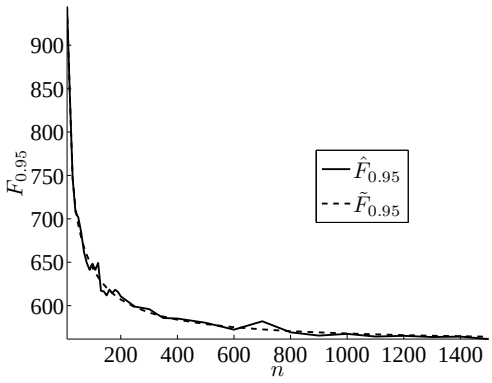


Рис. 2: Зависимость 0.95-квантиля $X^2$ от объёма выборки $n$ при $s = 1$, $v = 500$, $N = 1000$ и ее интерполяция.

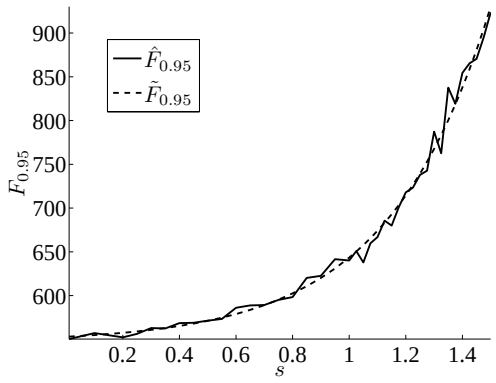


Рис. 3: Зависимость 0.95-квантиля $X^2$ от параметра $s$ при $n = 100$, $v = 500$, $N = 1000$ и ее интерполяция.

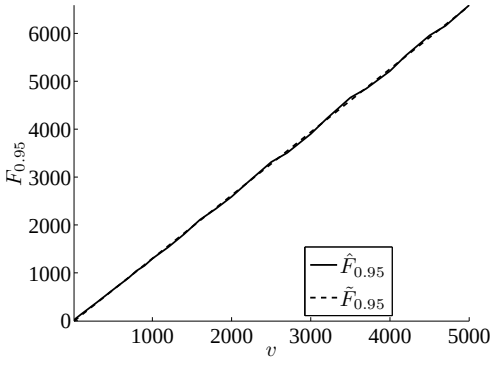


Рис. 4: Зависимость 0.95-квантиля $X^2$ от $v = |\Omega|$ при $s = 1$, $n = 100$, $N = 1000$ и ее интерполяция.

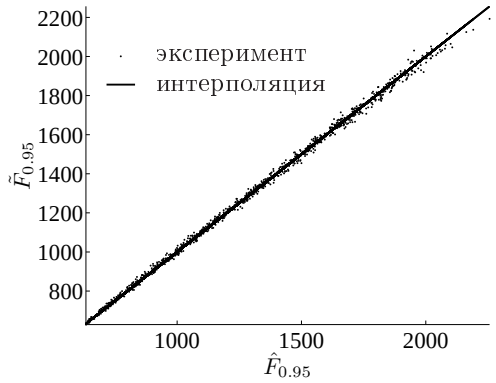


Рис. 5: Зависимость 0.95-квантилей, аппроксимированных моделью $\tilde{F}^4_{1-\alpha}$, от их эмпирических значений при различных $s$, $n$, $v$.

На рис. 3 показана зависимость 0.95-квантиля от показателя $s$ в законе Ципфа и её интерполяция функцией $\tilde{F}_{1-\alpha}(s) = F + GH^s$ с параметрами $F$, $G$, $H$.

На рис. 4 показана зависимости 0.95-квантиля от параметра $v = |\Omega|$ и её линейная интерполяция $\tilde{F}_{1-\alpha}(v) = I + Jv$ с параметрами $I$, $J$.

**Построение регрессионного теста.** Чтобы найти общий вид зависимости $\tilde{F}_{1-\alpha}(s, v, n)$, применяется эмпирический подход. Формируется обучающая выборка из 1000 троек $(s, v, n)$, равномерно выбранных из параллелепипеда $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$. Для каждой тройки вычисляется значение $\hat{F}_{n,0.95}$.

Для поиска нелинейной регрессионной зависимости используется алгоритм символьной регрессии MVR-composer [23, 24]. Преимущество этого алгоритма в том, что он автоматически подбирает формулу регрессии среди всевозможных суперпозиций заданного множества элементарных функций. В рассматриваемом случае

MVR-composer находит следующую модель регрессии:

$$\tilde{F}^1_{1-\alpha}(s,v,n) = (A + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4})(F + GH^s)(I + Jv)$$

и определяет оптимальные значения 10 параметров $A, B, C, D, E, F, G, H, I, J$. Рассматриваются также некоторые упрощения этой модели:

$$\tilde{F}^2_{1-\alpha}(s,v,n) = A(1 + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4})(1 + GH^s)(1 + Jv);$$
$$\tilde{F}^3_{1-\alpha}(s,v,n) = A(1 + Bn^{-c})(1 + GH^s)(1 + Jv);$$
$$\tilde{F}^4_{1-\alpha}(s,v,n) = Av(1 + Bn^{-c})(1 + GH^s);$$
$$\tilde{F}^5_{1-\alpha}(s,v,n) = Av(1 + GH^s);$$
$$\tilde{F}^6_{1-\alpha}(s,v,n) = Av(1 + Bn^{-c}).$$

Параметры этих моделей настраиваются с помощью функции nlinfit программы Matlab. Начальные приближения всех параметров полагаются равными 1, кроме параметра $A$, который инициализируется средним значением $\hat{F}_{n,1-\alpha}/v$ по всей выборке. Получаются следующие значения среднеквадратичной ошибки (СКО) на обучающей и контрольной выборках из 1000 случайных троек $(s,v,n)$ каждая:

| модель | $\tilde{F}^1_{1-\alpha}$ | $\tilde{F}^2_{1-\alpha}$ | $\tilde{F}^3_{1-\alpha}$ | $\tilde{F}^4_{1-\alpha}$ | $\tilde{F}^5_{1-\alpha}$ | $\tilde{F}^6_{1-\alpha}$ |
|---|---|---|---|---|---|---|
| число параметров | 10 | 8 | 6 | 5 | 3 | 3 |
| СКО (обучение) | 16.3 | 16.8 | 16.8 | 16.7 | 52.2 | 43.7 |
| СКО (контроль) | 15.8 | 16.1 | 16.0 | 16.0 | 50.9 | 43.8 |

Сравнение СКО на обучающей и контрольной выборках показывает, что переобучения нет ни в одной из моделей. Модель $\tilde{F}^4_{1-\alpha}$ представляется оптимальной по точности и числу параметров. Дальнейшее упрощение модели приводит к резкому увеличению СКО. Оптимальные значения параметров для неё: $A = 0.913$, $B = 3.98$, $c = 0.636$, $G = 0.00458$, $H = 36.8$.

На рис. 4 показан график зависимости 0.95-квантилей, аппроксимированных моделью $\tilde{F}^4_{0.95}$, от их эмпирических значений при различных $s$, $n$, $v$. Сплошной линией изображена «идеальная» прямая $\tilde{F} = \hat{F}$.

Таким образом, в отличие от классического критерия хи-квадрат, квантиль распределения статистики $X^2$ существенно зависит от объёма выборки $n$ и от вида распределения $p(x)$, в частности, от показателя степени $s$ в законе Ципфа, отвечающего за разреженность распределения. Построенная регрессионная модель довольно точно

описывает зависимость 0.95-квантили от параметров $s$, $n$, $v$. Эту зависимость можно построить один раз вместо того, чтобы строить тест для каждого распределения $p(x)$. Предварительно необходимо убедиться, что распределение $p(x)$ описывается законом Ципфа и найти значение параметра $s$. Данное обстоятельство сужает область применимости регрессионного теста.

**Анализ качества регрессионного теста.** В эксперименте оцениваются вероятности ошибок первого и второго рода предложенного регрессионного теста.

*Ошибкой первого рода* называется отклонение нулевой гипотезы при условии её истинности. Вероятность ошибки первого рода равна уровню значимости $\alpha = 0.05$. Для эксперимента генерируется контрольная выборка из 500 различных троек $(s, v, n)$, равномерно распределённых на параллелепипеде $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$. Для каждой тройки генерируется 1000 выборок объёма $n$ из распределения Ципфа $p(x)$ с параметрами $v$ и $s$ и вычисляется значение статистики $X^2$. Вероятность ошибки первого рода оценивается как доля выборок, для которых нулевая гипотеза отклоняется: $X^2 > \tilde{F}^4_{0.95}(s, v, n)$. Оценка вероятности ошибки первого рода составляет $0.0496 \pm 0.0141$ с доверительной вероятностью 0.95.

*Ошибкой второго рода* называется принятие гипотезы $H_0\colon p(x)$ при условии истинности ее альтернативы $H_1\colon p'(x)$. Вероятность ошибки второго рода существенно зависит от альтернативы — чем более похожи распределения $p(x)$ и $p'(x)$, тем больше вероятность ошибки. Исследуется способность теста различать распределения, отличающиеся на небольшом числе элементов $x$ из $\Omega$. Из множества $\Omega = \{1, \ldots, v\}$ выделяется подмножество элементов с наибольшими вероятностями: $\Omega_0 = \{x\colon p(x) > \mu p(1)\}$ при заданном $\mu \in (0, 1)$. Распределение $p'(x)$ строится из $p(x)$ следующим образом: выбираются $K$ различных случайных элементов множества $\Omega_0$ и их вероятности меняются местами с вероятностями $K$ различных случайных элементов множества $\Omega \backslash \Omega_0$.

Из полученного распределения $p'(x)$ генерируются выборки, для каждой строится эмпирическое распределение $\hat{p}(x)$ и вычисляется статистика $X^2$. Если $X^2 \leqslant \tilde{F}^4_{0.95}(s, v, n)$, то для данной выборки гипотеза $H_0$ ошибочно принимается. Доля выборок, при которых это происходит, является оценкой вероятности ошибки второго рода.

Для каждого $K$ генерируется 200 различных троек $(s, v, n)$ из равномерного распределения на параллелепипеде $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$ и вычисляется 200 оценок вероятности ошибки второго рода. На рис. 6 и рис. 7 показаны за-

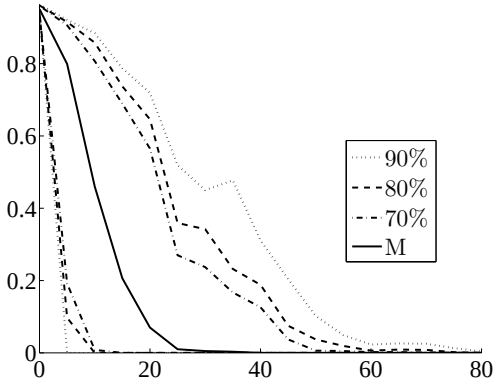


Рис. 6: Зависимость вероятности ошибки второго рода от $K$ при $\mu = 0.01$.

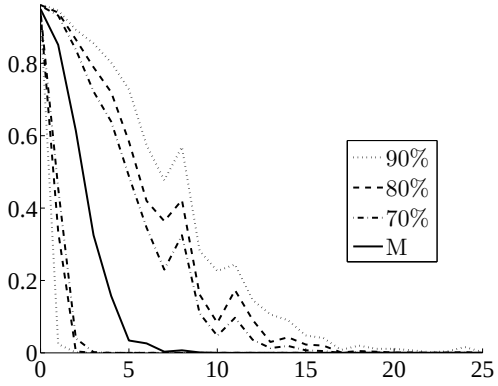


Рис. 7: Зависимость вероятности ошибки второго рода от $K$ при $\mu = 0.05$.

висимости медианы $M$ и доверительных границ 90%, 80%, 70% вероятности ошибки второго рода от числа перестановок $K$ при $\mu = 0.01$ и $\mu = 0.05$. По мере увеличения $K$ распределения $p(x)$ и $p'(x)$ все сильнее отличаются, и вероятность ошибки второго рода уменьшается. По мере увеличения $\mu$ различия становятся менее контрастными, и вероятность ошибки второго рода убывает медленнее. При $\mu = 0.01$ она становится меньше 0.1 при $K = 20$, при $\mu = 0.05$ она достигает этого значения при $K = 5$.

Отсюда, в частности, можно сделать вывод, что различные тексты, отличающиеся лишь 5 высокочастотными терминами, в среднем довольно надёжно различаются по их случайным фрагментам.

# 3 Проверка гипотезы условной независимости, основанная на статистике хи-квадрат

## 3.1 Вероятностные тематические модели

Чтобы оценить качество тематической модели, необходимо проверить, выполняется ли гипотеза условной независимости (1.2) — важнейшее базовое предположение модели (1.3) — для каждой пары документ–тема $(d, t)$ (проверка гипотезы для пар слово–тема $(w, t)$ выполняется аналогичным образом, отличия только в обозначениях). Тема $t$ описывается распределением $\hat{p}(w \mid t)$. Выборка слов документа $d$, относящихся к теме $t$, согласно модели, образует эмпирическое распределение $\hat{p}(w \mid d, t)$. Оба распределения оцениваются согласно (1.4) в процессе построения тематической модели. Чтобы проверить, действительно ли данная выборка могла быть получена из распределения $\hat{p}(w \mid t)$, используется критерий согласия, основанным на статистике

хи-квадрат (2.1):

$$X^2_{dt} = n_{dt} \sum_{w:\, n_{wt}>0} \frac{\big(\hat{p}(w \,|\, d,t) - \hat{p}(w \,|\, t)\big)^2}{\hat{p}(w \,|\, t)}. \tag{3.1}$$

Число различных слов в теме может быть намного больше, чем число слов в документе. Следовательно, распределения являются разреженными и к ним неприменим асимптотический критерий хи-квадрат. Поэтому статистические тесты строятся методом сэмплирования, для каждой темы $t \in T$ отдельно.

Экспериментально установлено, что для больших корпусов текстов на естественных языках закон Ципфа или более сложные параметрические законы (например Ципфа–Мандельброта) выполняются с неплохой точностью [19, 20]. Для ускорения проверки гипотезы условной независимости предлагается двухэтапный тест. Сначала проверяется согласие каждой темы $t$ с выбранным параметрическим законом. Если согласие есть, то строится один регрессионный тест для всех таких тем. Для каждой из остальных тем строится отдельный тест на основе сэмплирования.

## 3.2 Сэмплирование без возвращений

Проверки согласия документных эмпирических распределений $\hat{p}(w \,|\, d,t)$, $d \in D$ с распределением $\hat{p}(w \,|\, t)$, вообще говоря, не являются независимыми, поскольку имеется тождество, связывающие эти распределения друг с другом:

$$\hat{p}(w \,|\, t) = \sum_{d \in D} \hat{p}(w \,|\, d,t)\hat{p}(d \,|\, t). \tag{3.2}$$

Документы являются выборками без возвращений из распределения $\hat{p}(w \,|\, t)$, тогда как обычно критерии согласия предполагают выборку с возвращениями. Наличие дополнительного ограничения (3.2) может и не влиять на результаты тестов или влиять несущественно, особенно на коллекциях большого размера. Однако это лишь предположение, которое необходимо проверить. Для этого строится более точный тест на основе сэмплирования *без возвращений*, учитывающий, что последовательность слов, образующих тему $t$, разрезается на документы в пропорциях $\hat{p}(d \,|\, t)$.

**Построение теста сэмплированием без возвращений.** Рассматривается последовательность терминов длины $n_t$, образующая распределение $\hat{p}(w \,|\, t)$. Генерируется $N$ случайных перестановок этой последовательности. Каждая из полученных последовательностей $W_j$, $j = 1, \ldots, N$ делится на «документы» — подпоследовательности терминов $W_{jd}$ длины $n_{dt}$ каждая, $d \in D$. По каждому «документу» $W_{jd}$ строится эмпирическое распределение $\hat{p}_j(w \,|\, d,t)$ и вычисляется значение статистики хи-

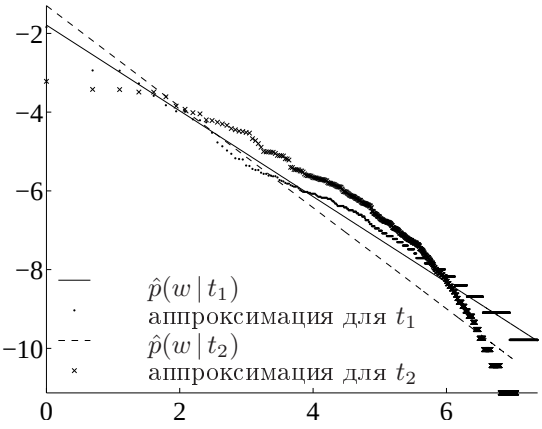


Рис. 8: Аппроксимация эмпирических распределений слов законом Ципфа (для двух тем, в логарифмических осях).

квадрат $X^2_{jd}$. Для каждого $d \in D$ по множеству значений статистики $X^2_{1d}, \ldots, X^2_{Nd}$ строится эмпирическая функция распределения $\hat{F}_d(X^2)$ и вычисляется её $(1-\alpha)$-квантиль $\hat{F}_{d,1-\alpha}$. Число $N$ рекомендуется брать не менее 1000 при типичном значении $\alpha = 0.05$.

Стоит отметить, что в тесте без возвращений квантиль строится для каждого документа $d$, тогда как тест с возвращениями строится для каждого значения длины документа $n$. Построение теста без возвращений более ресурсоёмко и требует $O(n_t N \log N)$ операций вместо $O(n_{\max} N \log N)$, где $n_{\max} = \max\limits_{d \in D} n_{td}$.

**Применение теста сэмплированием без возвращений.** Проверка гипотезы условной независимости для пары документ–тема $(d, t)$ заключается в вычислении статистики $X^2_{dt}$ по формуле (3.1) и проверке неравенства $X^2_{dt} > \hat{F}_{d,1-\alpha}$. Если оно выполнено, то гипотеза условной независимости отвергается для данной пары $(d, t)$.

## 3.3 Вычислительные эксперименты

Эксперименты проводились на коллекции из $|D| = 2000$ авторефератов диссертаций на русском языке. Мощность словаря после предварительной обработки данных (лемматизации и удаления стоп-слов) составляет $|W| = 20211$ слов, длина документов от 1000 до 4000 слов. Строились две тематические модели — PLSA [10] и LDA-GS [12, 13] с помощью алгоритма, описанного в [22]. Число тем $|T| = 100$.

**Выполняется ли закон Ципфа для тем?** На рис. 8 показаны графики эмпирических распределений и закона Ципфа для двух из 100 тем $t_1$ и $t_2$ в модели LDA, в логарифмических осях. По горизонтальной оси откладывается логарифм номера

слова, слова упорядочены по частоте. По вертикальной оси откладывается логарифм вероятности слова. Оптимальные значения параметра закона Ципфа: $s = 1.04$ для $t_1$, $s = 1.28$ для $t_2$. Хотя «на глаз» соответствие неплохое, особенно для $t_1$, нулевая гипотеза отклоняется для обоих тем. Более того, большинство тем согласуются с законом Ципфа лишь при крайне низких уровнях значимости, меньших 0.05. Это объясняется тем, что при выборках длины $n_t$ порядка $10^3$–$10^5$ критерии согласия чувствительны даже к незначительным различиям распределений, и одного параметра в законе Ципфа не достаточно для описания эмпирических распределений.

**Сравнение тестов без возвращений и с возвращениями.**

Для модели PLSA рассматривается одна тема из $|D_t| = 1992$ документов суммарной длины $n_t = 87026$ слов. В тестах без возвращений и с возвращениями нулевая гипотеза принимается для 1674 и 1688 документов соответственно. Решения отличаются на 22 документах из 1992. Оба теста дают примерно одинаковый результат: гипотеза условной независимости отклоняется для 15% документов.

Для модели LDA-GS рассматривается тема из $|D_t| = 1114$ документов суммарной длины $n_t = 63805$ слов. Нулевая гипотеза принимается для 1032 и 1035 документов соответственно. Решения отличаются на 7 документах из 1114. Оба теста снова дают примерно одинаковый результат: нулевая гипотеза отклоняется для 7% документов.

Таким образом, результаты тестов без возвращений и с возвращениями почти одинаковы, однако тест с возвращениями менее ресурсоёмкий.

**Определение оптимального числа тем.** Определение оптимального числа тем в коллекции — одно из возможных применений теста для проверки гипотезы условной независимости. Для модельной коллекции с числом документов $D = 500$ длин $n_d = 120$, мощностью словаря $|W| = 200$, параметрами распределения Дирихле $\alpha_0 = \beta_0 = 0.1$ и числом тем $T_0 = 10$ проводятся следующие эксперименты. Для каждого числа тем $T = 2, 3 \ldots 30$ строится тематическая модель LDA и выполняется проверка гипотезы условной независимости для каждой пары $(w, t)$ с помощью критерия, основанного на статистике хи-квадрат. Зависимость средней доли слов, не прошедших гипотезу на уровне значимости 0.05, от задаваемого числа тем изображена на рис. 9. Видно, что при числе тем, равном или большем оптимального $T = 10$, гипотеза условной независимости не отвергается.

Проверка гипотезы условной независимости с помощью критерия, основанного на статистике хи-квадрат, является довольно ресурсоемкой, т. к. требует построения

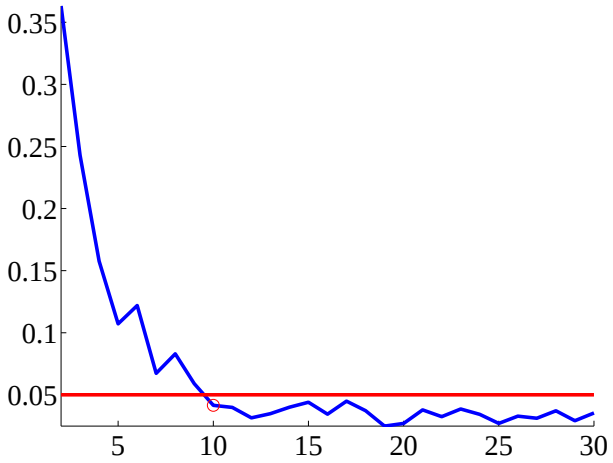


Рис. 9: Зависимость средней доли слов, не прошедших гипотезу, от числа тем.

эмпирической функции распределения статистики путем многочисленных сэмплирований. Далее предлагается другой, более эффективный подход к проверке данной гипотезы.

# 4 Критерии проверки независимости случайных величин

## 4.1 Точный тест Фишера.

Точный тест Фишера [18] используется для проверки отсутствия взаимосвязи между двумя переменными в таблице сопряженности размерности $2 \times 2$. При этом уровень значимости вычисляется при известных суммах по строкам и столбцам.

Пусть $X$ и $Y$ — две случайные величины, принимающие значения из множеств $\Omega_X = \{x_1, x_2\}$ и $\Omega_Y = \{y_1, y_2\}$ соответственно. Требуется проверить гипотезу о независимости $X$ и $Y$ по выборке длины $n$, состоящей из пар $(X_i, Y_i), \quad i = 1, 2 \ldots n$. Для этого составляется таблица сопряженности $F_{XY} = (a, b, c, d)$:

| | $y_1$ | $y_2$ | $\sum$ |
|---|---|---|---|
| $x_1$ | $a$ | $b$ | $a+b$ |
| $x_2$ | $c$ | $d$ | $c+d$ |
| $\sum$ | $a+c$ | $b+d$ | $a+b+c+d=n$ |

Число наблюдений вида $(x_1, y_1)$ обозначается $a$, вида $(x_1, y_2)$ — $b$, $(x_2, y_1)$ и $(x_2, y_2)$ —

$c$ и $d$ соответственно. Вероятность получить данную таблицу сопряженности при условии истинности нулевой гипотезы задается гипергеометрическим распределением:

$$P(F_{XY}) = \frac{C_{a+b}^{a} C_{c+d}^{c}}{C_{n}^{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}, \tag{4.1}$$

где $C_n^k$—биномиальный коэффициент, $(a+b)$ и $(c+d)$ — маргинальные суммы по строкам таблицы, $(a+c)$ и $(b+d)$ — по столбцам. Уровень значимости для двустороннего варианта теста Фишера:

$$\text{Pvalue} = \sum_{G_{XY}: P(G_{XY}) \leqslant P(F_{XY})} P(G_{XY}),$$

где $G_{XY}$ — всевозможные таблицы сопряженности с той же суммой элементов в каждой строке и каждом столбце, что и в таблице $F_{XY}$. Для проверки гипотезы против односторонней альтернативы о том, что наблюдение $x_1$ чаще встречается в совокупности с наблюдением $y_1$, чем с $y_2$, уровень значимости вычисляется по формуле:

$$\text{Pvalue} = \sum_{G_{XY}} P(G_{XY}), \tag{4.2}$$

где $G_{XY} = (a', b', c', d')$ — таблицы сопряженности с теми же маргинальными суммами, что и $F_{XY}$, но обладающие не меньшими элементами первой строки первого столбца, чем $F_{XY}$:

$$\begin{cases} a' \geqslant a \\ a' + b' = a + b \\ c' + d' = c + d \\ a' + c' = a + c \end{cases} \tag{4.3}$$

Условие $b' + d' = b + d$ следует из остальных. В случае больших выборок используется тест хи-квадрат, но он неприменим, когда математические ожидания значений в любой из ячеек таблицы с заданными границами ниже 5 [21]. Дело в том, что приближение выборочного распределения испытуемой статистической величины распределением хи-квадрат оказывается неадекватным при несбалансированном распределении данных среди ячеек таблицы, а также при малых размерах выборки. В таких случаях применяется точный тест Фишера, который не зависит от особенностей выборки.

## 4.2 Множественное использование точного теста Фишера

В более общем случае случайные величины $X$ и $Y$ могут принимать значения из конечных множеств любых размеров $\Omega_X = \{x_1, x_2 \dots x_{k_X}\}$ и $\Omega_Y = \{y_1, y_2 \dots y_{k_Y}\}$. Тест Фишера может быть применен и для анализа таблиц сопряженности размера $k_X \times k_Y$ [25], но он становится трудновычислимым при $k_X \gg 1$, $k_Y \gg 1$, т. к. требует перебора всевозможных таблиц с фиксированными маргинальными суммами. Для проверки гипотезы независимости всех документов от всех слов в теме линейные размеры таких таблиц достигают нескольких десятков тысяч, поэтому данный подход оказывается неприемлемым.

В работе для каждой пары $(i, j)$, где $i = 1, 2 \dots k_X$, $j = 1, 2 \dots k_Y$ составляется таблица сопряженности $F_{X_i Y_j}$:

| | $y_j$ | $\Omega_Y \backslash y_j$ | $\sum$ |
|---|---|---|---|
| $x_i$ | $a_{ij}$ | $b_{ij}$ | $a_{ij} + b_{ij}$ |
| $\Omega_X \backslash x_i$ | $c_{ij}$ | $d_{ij}$ | $c_{ij} + d_{ij}$ |
| $\sum$ | $a_{ij} + c_{ij}$ | $b_{ij} + d_{ij}$ | $a_{ij} + b_{ij} + c_{ij} + d_{ij} = n$ |

Далее вычисляются уровни значимостей $\mathrm{Pvalue}_{X_i Y_j}$ $\forall (i, j) : i = 1, 2 \dots k_X$, $j = 1, 2 \dots k_Y$ по формуле (4.2) (рассматривается односторонняя альтернатива). При условии истинности нулевой гипотезы распределение вычисленных уровней значимости $\{\mathrm{Pvalue}_{X_i Y_j}\}$ должно быть близко к равномерному и с вероятностью $\alpha$ (типичное значение $\alpha = 0.05$) должно выполняться $\mathrm{Pvalue}_{X_i Y_j} < \alpha$. Пусть $\mathrm{Pvalue}_{X_i Y_j} < \alpha$ для $k_{XY}^{\alpha}$ из $k_{XY}$ пар $(i, j)$. С помощью биномиального теста проверяется гипотеза о том, что $\mathrm{Pvalue}_{X_i Y_j} < \alpha$ с вероятностью, равной $\alpha$, против односторонней альтернативы о том, что вероятность больше $\alpha$. Достигаемый уровень значимости:

$$\mathrm{Pvalue}_{XY} = \sum_{i=k_{XY}^{\alpha}}^{k_{XY}} C_{k_{XY}}^{i} \, \alpha^i \, (1 - \alpha)^{k_{XY} - i}. \tag{4.4}$$

При $\mathrm{Pvalue}_{XY} < \alpha$ гипотеза о независимости $X$ и $Y$ отвергается.

**Алгоритм 5.1.** Проверка гипотезы условной независимости для темы $t$

**Вход:** $t$, $n_t$, $\alpha$, $n_{dwt}$, $n_{dt}$, $n_{wt}\, \forall d \in D, w \in W$;

**Выход:** отвергнуть или нет гипотезу на уровне значимости $\alpha$;

**для всех** документов $d \in D$
  **для всех** слов $w \in W$
    **если** $n_{dt} > 0$ **и** $n_{wt} > 0$ **то**
      проверить независимость $d$ и $w$ в теме $t$:
      составить таблицу сопряженности:

| | $w$ | $W \backslash w$ |
|---|---|---|
| $d$ | $n_{dwt}$ | $n_{dt} - n_{dwt}$ |
| $D \backslash d$ | $n_{wt} - n_{dwt}$ | $n_t - n_{dt} - n_{wt} + n_{dwt}$ |

      провести точный тест Фишера и вычислить значение уровня значимости: $\text{Pvalue}_{dwt}$
провести биномиальный тест:
вычислить значение достигаемого уровня значимости $P_t$
**если** $P_t < \alpha$ **то**
  гипотеза условной независимости отвергается
**иначе**
  данные гипотезе не противоречат

# 5 Проверка гипотезы условной независимости с помощью точного теста Фишера

## 5.1 Применение точного теста Фишера для проверки гипотезы условной независимости

**Проверка гипотезы условной независимости для темы** $t$. Для того, чтобы определить, описывает ли тематическая модель данную тему $t$, требуется проверить гипотезу о независимости документов и слов в теме, т. е. двух случайные величин, принимающих значения из множеств $D_t = \{d \in D : n_{dt} > 0\}$ и $W_t = \{w \in W : n_{wt} > 0\}$ соответственно. Для этого предлагается алгоритм, описанный в предыдущем разделе, где $X$—документ, $Y$—слово в теме $t$ (алгоритм 5.1).

**Проверка гипотезы условной независимости для пар** $(w, t)$ **и** $(d, t)$ . Проверку гипотезы условной независимости для пары слово–тема (документ–тема) пред-

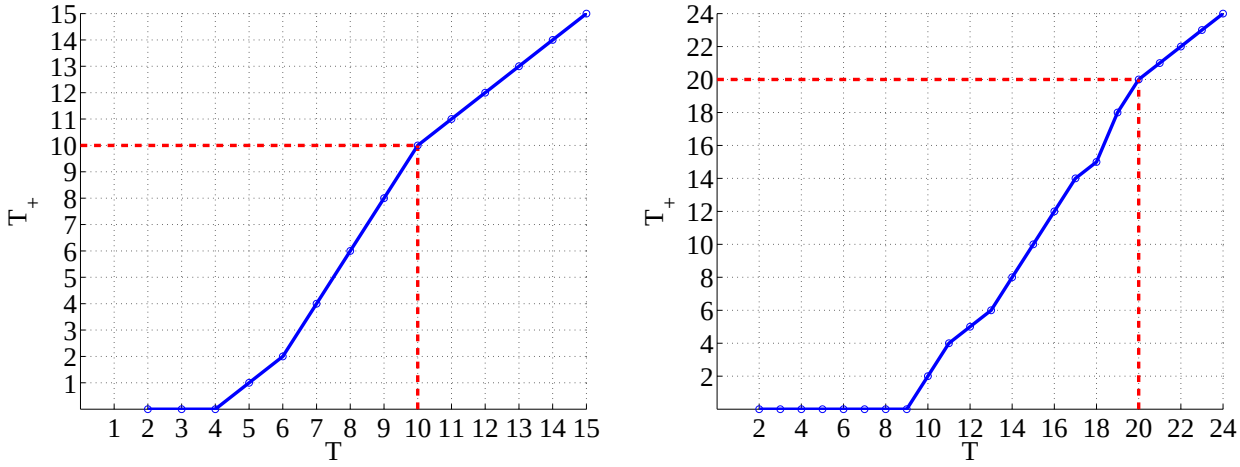


Рис. 10: Зависимость числа тем, для которых гипотеза не отвергается, от задаваемого числа тем при оптимальном числе тем $T = 10$ (слева) и $T = 20$ (справа).

лагается проводить аналогичным образом: рассматриваются случайные величины, принимающие значения из множеств: $\{w, W\backslash w\}$ и $\{d \in D : n_{dt} > 0\}$ (для пары документ–тема $\{d, D\backslash d\}$ и $\{w \in W : n_{wt} > 0\}$) и проверяется их независимость по алгоритму, основанному на множественном использовании точного теста Фишера.

## 5.2 Вычислительные эксперименты

Рассматриваются две модельные коллекции: в первой коллекции число документов $D = 500$, мощность словаря $|W| = 200$, длина документов $n_d = 120$, количество тем $T_0 = 10$, параметры распределения Дирихле $\alpha_0 = \beta_0 = 0.1$; во второй коллекции — $D = 900$, $|W| = 300$, $n_d = 120$, $T_0 = 10$, $\alpha_0 = \beta_0 = 0.1$. Для каждой коллекции строятся тематические модели LDA с числом тем $T = 2, 3 \ldots 15$ и $T = 2, 3 \ldots 24$ соответственно, и для каждой темы выполняется проверка гипотезы условной независимости. На рис. 10 показаны зависимости числа тем, для которых гипотеза условной независимости не отвергается $T_+$, от задаваемого числа тем $T$ для двух коллекций. При числе тем, большем либо равного оптимального, гипотеза условной независимости принимается для всех тем.

# 6 Выводы

Предложены критерии согласия на основе сэмплирования для разреженных дискретных распределений, выходящих за границы применимости классических асимп-

тотических критериев. Предложен рекуррентный алгоритм построения теста на основе сэмплирования. Для параметрического случая, когда проверяется согласие эмпирических данных с распределением Ципфа, построен регрессионный тест, подобрана модель регрессии и проведён анализ ошибок первого и второго рода. Также предложен тест для проверки независимости двух случайных величин на основе множественного использования теста Фишера для таблиц сопряженности $2 \times 2$. Рассмотрено применение предложенных тестов для проверки адекватности вероятностных тематических моделей.

# Список литературы

[1] L. Fei-Fei and P. Perona *A Bayesian hierarchical model for learning natural scene categories*, IEEE Computer Vision and Pattern Recognition, pages 524-531, 2005.

[2] D.Blei and M.Jordan *Modelling annotated data* In Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 127-134. ACM Press, 2003.

[3] E.Bart, M.Welling, and P. Perona *Unsupervised organization of image collections: Taxonomies and beyond*, Transactions on Pattern Recognition and Machine Intelligence, 2010.

[4] J.Li, C.Wang, Y.Lim, D.Blei, and L.Fei-Fei *Building and using a semantivisual image hierarchy*, In Computer Vision and Pattern Recognition, 2010.

[5] S.Deerwester; S. T. Dumais; T. K. Landauer; G. W. Furnas and R. A. Harshman *Indexing by latent semantic analysis*, Journal of the Society for Information Science, 41(6), 391-407, 1990.

[6] Papadimitriou, Christos; Raghavan, Prabhakar; Tamaki, Hisao; Vempala, Santosh *Latent Semantic Indexing: A probabilistic analysis*, Proceedings of ACM PODS, 1998.

[7] von Davier M. *Bootstrapping goodness-of-fit statistics for sparse categorical data-results of a monte carlo study*, Methods of Psychological Research Online,1997.

[8] Zelterman D. *Goodness-of-fit tests for large sparse multinomial distributions*, Journal of the American Statistical Association, Pp. 624-629, 1987.

[9] Daud, Ali and Li, Juanzi and Zhou, Lizhu and Muhammad, Faqir *Knowledge discovery through directed probabilistic topic models: a survey*, Frontiers of Computer Science in China,Pp. 280-301, 2010.

[10] Hofmann, Thomas *Probabilistic latent semantic indexing*,Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Pp.50-57,1999.

[11] Dempster A.P., Laird N. M., Rubin D. B. *Maximum likelihood from incomplete data via the EM algorithm*, J. of the Royal Statistical Society, Series B., no.34, Pp.1-38, 1977.

[12] Blei, David M. and Ng, Andrew Y. and Jordan, Michael I. *Latent Dirichlet allocation*, Journal of Machine Learning Research, Pp.993-1022, 2003.

[13] Steyvers, Mark and Griffiths, Tom, *Finding scientific topics* ,Proceedings of the National Academy of Sciences, Pp.5228-5235, 2004.

[14] A. Asuncion and M. Welling and P. Smyth and Y. W. Teh, *On Smoothing and Inference for Topic Models*, Proceedings of the International Conference on Uncertainty in Artificial Intelligence, 2009.

[15] Yi Wang, *Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details*,2008.

[16] G.Golub and C.Reinsch, *Handbook for matrix computation II, Linear Algebra*, Springer-Verlag, New York, 1971.

[17] Mimno D., Blei D., *Bayesian checking for topic models*, 11th Conference on Empirical Methods in Natural Language Processing.— Association for Computantional Linguistics, 2011.—Pp. 227-237.

[18] Fisher, R. A. *On the interpretation of $\chi^2$ from contingency tables, and the calculation of P*, Journal of the Royal Statistical Society,1922 85(1):87-94.

[19] Бриллюэн Л. *Наука и теория информации*, М.: «Государственное издательство физико-математической литературы», 1960.—391с.

[20] Gelbukh A., Sidorov G. *Zipf and heaps laws' coefficients depend on language* //Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City, Lecture Notes in Computer Science.— No. 2004.— Springer-Verlag, 2001.—P. 332–335.

[21] A. L. Edwards *On the use and misuse of the chi-square test –– the case of the 2x2 contingency table*, Psychological Bulletin, Vol 47(4), Jul 1950, pp. 341-346.

[22] Воронцов К. В., Потапенко А. А. *Регуляризация, робастность и разреженность вероятностных тематических моделей*, Компьютерные исследования и моделирование.— 2012.—Т.4, №4.—стр. 693–706.

[23] Strijov V. *Search for a parametric regression model in an inductive-generated set*, Computational technologies, 2007 Vol. 12, no. 1, Pp. 93–102.

[24] Strijov V. *MVR Composer*, 2012 http://strijov.com/?p=84.

[25] Arthur W. Ghent *A Method for Exact Testing of 2X2, 2X3, 3X3, and Other Contingency Tables, Employing Binomial Coefficients*, American Midland Naturalist, Vol. 88, No. 1, Jul 1972, pp. 15-27.