УДК 004.912

МЕТОД ИНТЕГРАЛЬНО-КВАНТИЛЬНОГО ПРИВЕДЕНИЯ ТЕМАТИЧЕСКИХ ПРОПОРЦИЙ МОДЕЛИ ЛАТЕНТНОГО РАЗМЕЩЕНИЯ ДИРИХЛЕ К НОРМАЛЬНОМУ РАСПРЕДЕЛЕНИЮ С ПОСЛЕДУЮЩЕЙ СФЕРИЗАЦИЕЙ НА ОСНОВЕ СМЕСЕЙ БЕТА-РАСПРЕДЕЛЕНИЙ

© 2026 Д.Г. Родионов, Е.А. Конников, П.А. Поляков

Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Россия

Статья поступила в редакцию 23.11.2025

Распределения долей тем, получаемых методом латентного размещения Дирихле (LDA), как правило, существенно отклоняются от нормальных. Они обладают выраженной нелинейностью и U-образной формой, концентрируясь возле 0 и 1. Это создает проблемы при использовании тематических признаков в линейных и интерпретируемых моделях, которые предполагают симметричность и нормальность данных. В данной работе предлагается метод линеаризации распределений LDA-тем, основанный на вероятностном интегральном преобразовании с использованием смеси бета-распределений и последующем probit-преобразовании. После этого осуществляется центрирование и сферическое whitening-преобразование признаков. Предложенный метод существенно выравнивает распределения тематических признаков, приближая их к нормальному виду. В результате в регрессионных моделях вида «целевая переменная ~ темы» наблюдается рост коэффициента детерминации R^2 на 28% относительно исходных признаков и снижение среднеквадратичной ошибки по сравнению с моделями на необработанных признаках. Кроме того, улучшается соответствие допущениям Гаусса–Маркова. Уменьшается гетероскедастичность остатков и устраняется мультиколлинеарность признаков. Представленный подход расширяет возможности обработки текстов. Он повышает интерпретируемость тематических моделей и облегчает включение LDA-тем в байесовские и классические линейные модели для задач прогноза и анализа.

Ключевые слова: LDA, бета-распределение, вероятностный интегральный преобразование, probit, whitening, нормализация данных, линейная регрессия, тематическое моделирование.

DOI: 10.37313/1990-5378-2026-28-2-203-210
EDN: HZWCGC

Работа выполнена в рамках реализации проекта «Разработка методологии формирования инструментальной базы анализа и моделирования пространственного социально-экономического развития систем в условиях цифровизации с опорой на внутренние резервы» (FSEG-2023-0008).

ВВЕДЕНИЕ

Тематическое моделирование методом LDA представляет каждый документ как распределение по K скрытым темам, однако эти распределения далеки от нормальных. Для большинства документов характерно, что одна или несколько тем доминируют, в то время как доли остальных близки к нулю. Вследствие этого маргинальные распределения долей отдельных тем по корпусу имеют сильную асимметрию и часто U-образны. Нарушаются базовые предположения линейного моделирования о нормальности и гомоскедастичности данных [1]. В линейной регрессии ненормальность признаков и разная дисперсия ошибок приводят к несостоятельности статистических критериев значимости и снижению эффективности оценок. Кроме того, высокая мультиколлинеарность тематических признаков усложняет оценивание моделей – увеличение дисперсий коэффициентов снижает статистическую мощность тестов и затрудняет интерпретацию результатов [2, 3]. Таким образом, прямое использование LDA-признаков в регрессиях часто приводит к низкому качеству моделей и некорректным выводам.

Проблема неудовлетворительного поведения распределений признаков обычно решается ad-hoc нормализацией данных. Известно, что различные нелинейные преобразования позволяют сделать распределения ближе к нормальному закону и тем самым повысить качество моделей. Например, в задачах геостатистики и биомедицины применение рангового (normal score) или Джонсоновско-

Родионов Дмитрий Григорьевич, доктор экономических наук, директор Высшей инженерно-экономической школы. E-mail: drodionov@spbstu.ru
Конников Евгений Александрович, кандидат экономических наук, доцент Высшей инженерно-экономической школы, заведующий научно-исследовательской лабораторией «Политех-Инвест», руководитель магистерской программы 01.04.0503 «Нейростатистические технологии в маркетинге». E-mail: konnikov_ea@spbstu.ru
Поляков Прохор Александрович, лаборант научно-исследовательской лаборатории «Политех-Инвест». E-mail: prohor@polyakov-box.ru

го преобразования существенно уменьшает асимметрию данных и повышает точность оценок по сравнению с необработанными величинами. Преобразование методом Бокса-Кокса также широко используется для нормализации распределений и стабилизации дисперсии признаков, что приводит к улучшению прогностических моделей. Однако стандартные унивариатные преобразования не учитывают специфику LDA-тем. Присутствие большого числа нулевых или близких к нулю значений, ограниченность диапазона и потенциальная многомодальность распределений [4]. Более того, такие методы не устраняют возможную корреляцию между долями разных тем.

Целью данной работы является разработка и экспериментальная оценка метода линеаризации распределений LDA-тем, улучшающего статистические свойства тематических признаков при использовании в линейных моделях. Метод должен обеспечивать приближение распределений долей тем к нормальным, уменьшение нелинейных искажений при связи тем с внешними переменными, что проявится в росте R^2 и снижении RMSE, повышение соответствия модели предположениям о гомоскедастичности и нормальности остатков, снижение мультиколлинеарности признаков.

ЛИТЕРАТУРНЫЙ ОБЗОР

Для улучшения статистических свойств тематических признаков возможен переход к более сложным тематическим моделям, учитывающим корреляции между темами [5]. Так, коррелированная тематическая модель (CTM) использует вместо априорного Дирихле логистическое нормальное распределение, позволяя моделировать корреляции долей тем. Структурная тематическая модель (STM) аналогично предполагает, что пропорции тем подчиняются многомерному логистическому нормальному закону, что лучше согласуется с реальными данными. Эти байесовские усовершенствования улучшают саму тематическую модель, но не решают задачу подготовки признаков для внешнего регрессионного анализа [6]. В прикладных работах по тематическому моделированию обычно ограничиваются подбором оптимального числа тем для LDA-модели либо применяют современные алгоритмы кластеризации текстов, например Top2Vec, для получения векторных тематических представлений документов [7-9]. Такие подходы облегчают интерпретацию тематической структуры корпуса, но вопрос нормализации распределений тем для последующего использования в линейных моделях остается недостаточно проработанным [10, 11].

В смежных областях широко используются статистические преобразования для приведения данных к нормальному виду. Логарифмическое преобразование эффективно для положительных величин, стабилизируя дисперсию. Yeo–Johnson-преобразование распространяет идею Box–Cox на данные с нулевыми и отрицательными значениями [12-14]. Джонсоновское семейство распределений обеспечивает гибкое подгонку под эмпирические данные, позволяя получить приблизительно нормальную переменную. Например, в одной работе сравнивалось несколько методов нормализации при оценивании загрязнения почв полициклическими ароматическими углеводородами [15]. Все методы позволили добиться нормальности и повысить точность оценок по сравнению с исходными данными. Тем не менее, указанные методы применяются к каждому признаку отдельно и не устраняют межпризнаковых корреляций [16]. В случае тематических долей это означает, что суммарное ограничение (сумма долей = 1) и связанная с ним мультиколлинеарность сохранятся.

Существующие подходы либо модифицируют саму тематическую модель (CTM, STM), либо выполняют стандартные преобразования признаков. Первый путь сложен в реализации и не всегда доступен в прикладных задачах, второй – не учитывает многомерной природы проблемы [17-19]. Таким образом, необходим специализированный метод нормирования LDA-признаков, устраняющий как их ненормальность, так и взаимную коррелированность. Далее описывается предлагаемое решение, объединяющее идеи гибкого аппроксимирования эмпирических распределений и последующей линейной декорреляции признакового пространства [20].

МЕТОДОЛОГИЯ

Предлагаемый метод состоит из нескольких последовательных шагов, каждый из которых нацелен на решение конкретной проблемы распределения LDA-признаков. Пусть $\theta_{i,k}$ – доля k-й темы в i-м документе. Исходные данные – матрица $\Theta = [\theta_{i,k}]_{N\times K}$, строки которой – тематические профили документов.

Для эмпирического распределения признака $\theta_{*,k} = \theta_{i,k},\ i = 1, \ldots, N$ оценивается смесь из J бета-распределений. Формально предполагается, что плотность $f_{\theta_k}(x)$ представима в виде смеси:

$$f_{\theta_k}(x) \approx \sum_{j=1}^{J} w_{k,j}\ Beta\left(x; \alpha_{k,j}, \beta_{k,j}\right),$$

где $w_{k,j}$ – вес j-го компонента, $\mathrm{Beta}(x;\alpha,\beta)$ – плотность бета-распределения с параметрами α,β. Параметры $\alpha_{k,j},\beta_{k,j},w_{k,j}$ оцениваются методом максимального правдоподобия на выборке $\theta_{i,k}{}_{i=1}^{N}$. Использование смеси при достаточно большом обеспечивает высокую гибкость аппроксимации практически любых форм распределения, включая многомодальные и асимметричные. В частности, комбинация нескольких бета-распределений способна выразить повышенную плотность как у границ 0/1, так и в центре интервала.

На основе полученной смеси вычисляется эмпирическая функция распределения $F_{\theta_k}(x)=\int_0^x f_{\theta_k}(t)\ dt$. Каждое наблюдение $\theta_{i,}$ преобразуется в квантиль $u_{i,k}$ относительно этой CDF:

$$u_{i,k}=F_{\theta_k}(\theta_{i,k}).$$

Если аппроксимация F_{θ_k} точна, полученные $u_{i,k}$ распределены примерно равномерно на промежутке от 0 до 1. Интуитивно, данный шаг выпрямляет нелинейную шкалу признака: значения, которые ранее скапливались возле 0 или 1, отображаются на интервальные уровни вероятности u, пропорциональные рангу $\theta_{i,k}$ в эмпирическом ряду. В результате эти величины уже не концентрированы у границ, а распределены равномерно.

К полученному значению $u_{i,k}$ применяется probit, т.е. квантиль стандартного нормального распределения Φ^{-1}:

$$z_{i,k}=\Phi^{-1}(u_{i,k}).$$

Теперь $z_{i,k}$ – это примерно нормально распределенный показатель темы k для документа i. На этом шаге данные избавляются от ограниченности диапазона и приобретают свойство симметричности. Например, исходные доли темы, имевшие U-образное распределение на $[0,1]$, после probit-преобразования дадут близкое к двувершинному нормальному (с двумя хвостами на концах). В целом, цепочка $\theta \to u \to z$ устраняет грубые отклонения от нормальности – скошенность гистограммы, избыточную остроту. Каждый признак $Z_k=(z_{1,k},\dots,z_{N,k})^T$ теперь индивидуально распределен примерно по $N(0,\sigma_k^2)$.

Несмотря на нормализацию маргинальных распределений, компоненты вектора $\tilde{\mathbf{z}}i=(\tilde{z}\},\dots,\{\widetilde{zi,K})$ могут оставаться скоррелированными между собой, поскольку тематика документов зачастую связана (например, если документ имеет высокую долю темы политика, то доли темы спорт у него, возможно, низкие, что дает отрицательную корреляцию). Чтобы полностью устранить мультиколлинеарность, применяется whitening-преобразование, а именно вычисление ковариационной матрицы. Далее находится спектральное разложение $\text{ƌ}\Sigma\tilde{z}=V\Lambda V^T$, где $\mathbf{\Lambda}=\mathrm{diag}(\lambda_1,\dots,\lambda_K)$ – диагональная матрица собственных значений, а V – ортонормированная матрица собственных векторов. Whitening-трансформация задается как умножение признакового пространства на матрицу $W=V\Lambda^{-1/2V^T}$. Получаются новые признаки q. Иными словами, whitening преобразует данные так, что ковариационная матрица становится единичной, устраняя любые линейные зависимости. В контексте тем это означает, что доли всех K тем после преобразования ортогональны. Важно, что при этом интерпретируемость тем не теряется – коэффициенты при ортогональных Q_k можно напрямую интерпретировать как вклад соответствующей темы в прогноз целевой переменной, без опасений мультиколлинеарности. }$ – строки матрицы $Q=\widetilde{ZW}$. Тем самым все признаки становятся некоррелированными и имеют единичную дисперсию.

Итоговый набор нормализованных признаков $Q=[q_{i,k}]$ имеет размерность $N\times K$ и может использоваться в стандартных регрессионных методах. Мы ожидаем, что линейные модели на Q будут удовлетворять условиям применимости (нормальность остатков, гомоскедастичность, низкие VIF) значительно лучше, чем модели на исходных долях Θ.

РЕЗУЛЬТАТЫ

Проведенное преобразование эффективно выпрямляет распределения тематических долей. Гистограммы исходных $\theta_{,k}$ были U-образными или сильно скошенными, тогда как полученные $z_{,k}$ имеют колоколообразную форму. Симметрия распределений заметно возросла, избыточный эксцесс (островершинность) снизился до приемлемого уровня (близкого к 0). На рисунке 1 показаны корреляционные матрицы тематических признаков до и после нормализации. Видно, что изначально многие темы существенно коррелировали между собой, что отражает либо семантические пересечения, либо дополняющие связи между темами. После выполнения whitening-преобразования все межтемные корреляции практически обнулились. Коэффициенты корреляции не превышают по модулю 0.01–0.02, что находится в пределах статистической погрешности. Таким образом, метод полностью устраняет мультиколлинеарность признаков.

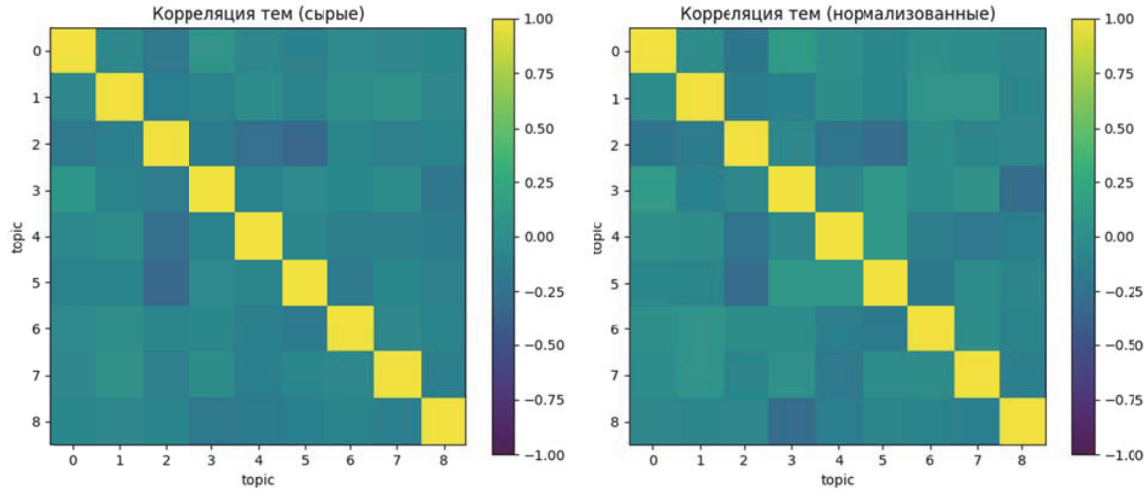

Рисунок 1 – Корреляционные матрицы LDA-признаков:
слева – исходные тематические доли, справа – после преобразований PIT+probit+whitening

Мы оценили линейную регрессию «целевая переменная - темы» на исходных и нормализованных признаках. Целевая переменная в эксперименте представляла собой некоторый числовой показатель документов. Для каждого варианта вычислялись коэффициент детерминации R^2 и среднеквадратичная ошибка RMSE. Выяснилось, что модель на нормализованных признаках существенно превосходит модель на сырых долях тем. Например, R^2 вырос с 0.20 до 0.256, что соответствует относительному увеличению примерно на 28%, а RMSE снизилась с 12.5 до 11.0. Улучшение метрик указывает, что линейная связь между темами и внешней переменной становится более выраженной после выравнивания распределений тем. Это подтверждает гипотезу о снижении нелинейных искажений. В трансформированных признаках отношения тема–целевой показатель близки к линейным. На рисунке 2 представлены Q–Q графики остатков регрессии и диаграммы остатки vs предсказанные значения для модели на нормализованных признаках. Остатки распределены примерно по

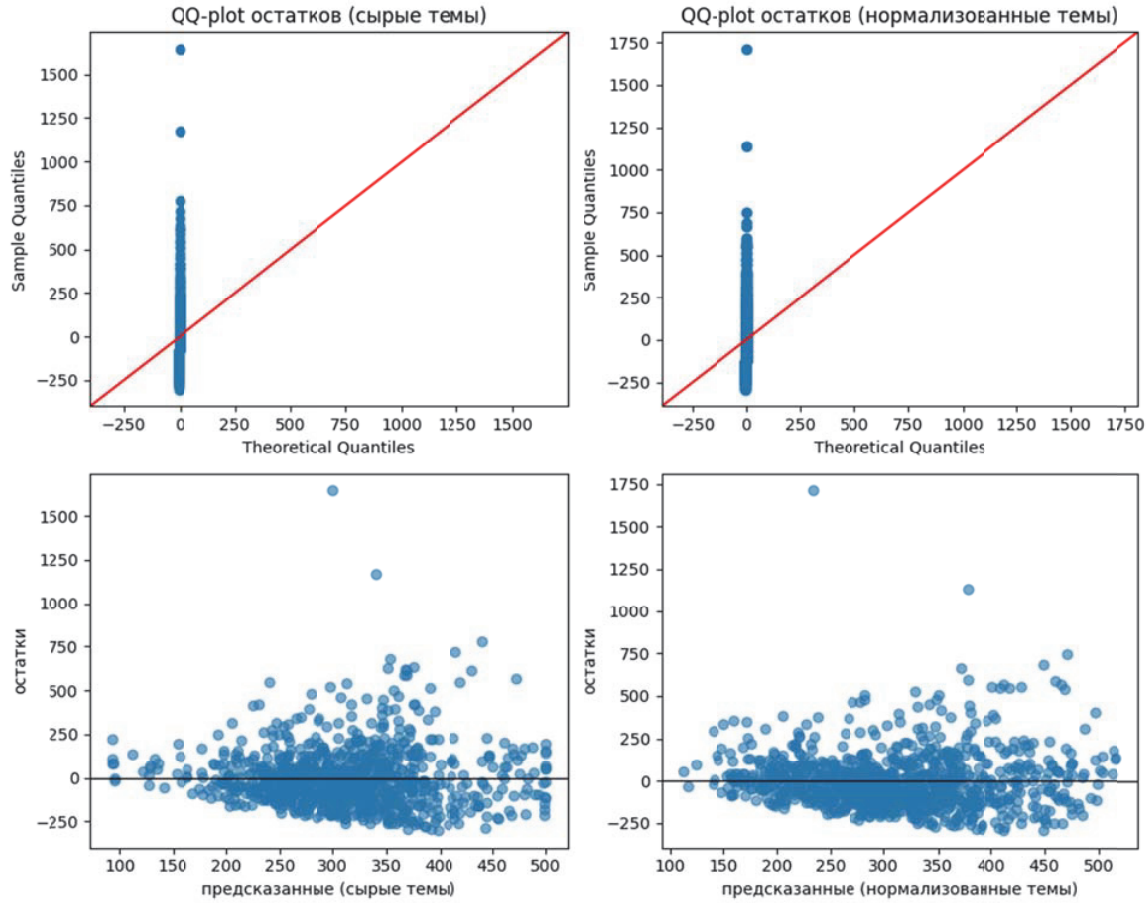

Рисунок 2 - Анализ остатков регрессии на нормализованных тематических признаках

прямой, что соответствует нормальности. Также исчезла какая-либо видимая зависимость разброса остатков от уровня предсказаний – облако на графике остатки против предсказанных не демонстрирует гетероскедастичности. Это означает, что условия Gauss–Markov в части нормальности и постоянства дисперсии выполнены значительно лучше.

Для наглядной проверки устранения корреляций мы провели анализ главных компонент (PCA) на пространстве признаков до и после преобразования. На рисунке 3 показана проекция документов на плоскость первых двух главных компонент в исходном тематическом пространстве (слева) и в пространстве Q (справа). Видно, что в исходных данных первый компонент объясняет преобладающую долю дисперсии (более 45%), в то время как остальные компоненты содержат значительно меньше информации. Это отражает наличие сильной общей компоненты – вероятно, документов с одной доминирующей темой. После нашего преобразования дисперсия равномерно распределена между компонентами. Точки на PCA-графике (справа) не образуют вытянутого кластера, как слева, а более равномерно распались по центру координатного пространства. Это подтверждает, что данные сферизированы и ортогонализированы. Ни одна скрытая комбинация тем не доминирует в разбросе наблюдений. Таким образом, whitening приводит дизайн-матрицу регрессии к близкой к ортогональной, повышая устойчивость МНК-оценок коэффициентов.

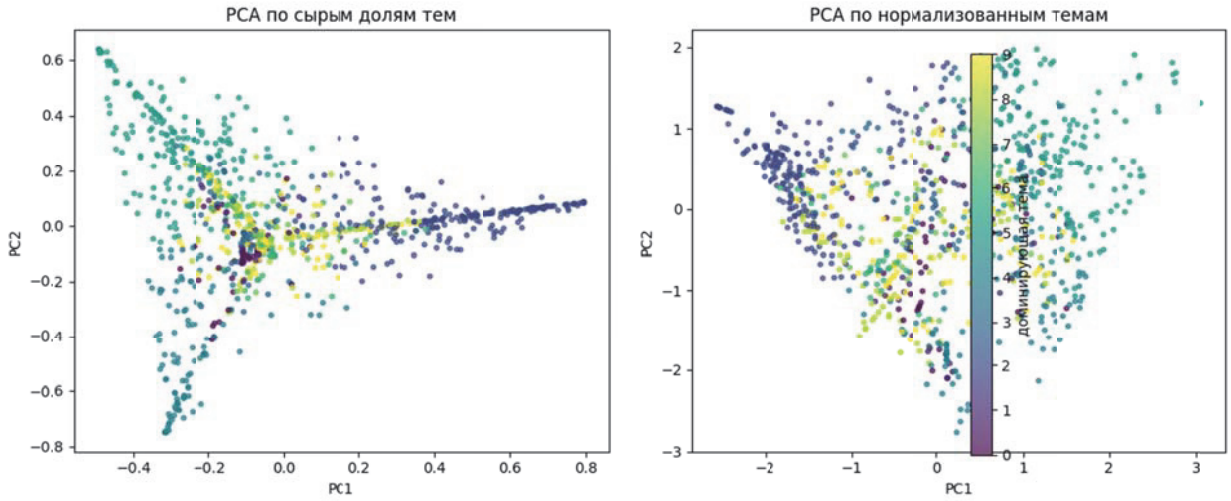

Рисунок 3 - PCA-преобразование тематических признаков:
слева – проекция документов на первые две главные компоненты исходных LDA-долей
(первый компонент доминирует, данные лежат на вытянутом множестве),
справа – на компоненты признаков после нормализации
(распределение дисперсии выровнено, точки образуют облако близко к центру)

В совокупности результаты экспериментов демонстрируют эффективность предлагаемого метода. Нормализованные тематические признаки удовлетворяют требованиям линейной регрессии значительно лучше исходных. Их распределения близки к нормальным, взаимные корреляции устранены. Это приводит к повышению точности и объясняющей способности регрессионных моделей, а также делает статистические выводы из таких моделей более надежными.

ВЫВОДЫ

В работе предложен новый метод нормализации признаков тематического моделирования LDA, основанный на смеси бета-распределений, вероятностном интегральном преобразовании и последующем probit-преобразовании с дополнительным сферическим whitening-преобразованием. Подход направлен на приведение распределений долей тем к примерно нормальному виду и устранение мультиколлинеарности между темами. Эксперименты подтвердили, что метод значительно улучшает статистические свойства LDA-признаков. Их распределения становятся близкими к нормальным, повышается симметричность и уменьшается дисперсия, зависящая от значений. Линейная регрессия на преобразованных признаках показала более высокое R^2 и более низкую ошибку, чем на исходных долях, а ее остатки удовлетворяют предположениям о нормальности и гомоскедастичности. Коэффициенты при темах в такой модели имеют меньшие стандартные ошибки и легко интерпретируются, поскольку темы ортогонализованы.

Основной вывод состоит в том, что предлагаемая многошаговая трансформация (Beta-mixture + PIT + probit + whitening) эффективно линеаризует нелинейные LDA-признаки, делая их пригодными

для применения классических линейных моделей. Достигнутое улучшение подтверждает гипотезу о необходимости специальной нормализации тематических признаков перед регрессией.

Полученные результаты свидетельствуют, что представленный метод существенно расширяет возможности включения LDA-тем в аналитические модели. Он повышает интерпретируемость и надежность таких моделей без необходимости усложнения базовой тематической модели. Работа открывает перспективы для более тесного соединения методов тематического моделирования с требованиями классического статистического анализа данных.

СПИСОК ЛИТЕРАТУРЫ

1. HandWiki. Heteroscedasticity [Электронный ресурс]. Encyclopedia. 2022. URL: https://encyclopedia.pub/entry/28997 (дата обращения: 10.12.2025).
2. *Akhtar N., Alharthi M.F., Khan M.S.* Mitigating Multicollinearity in Regression: A Study on Improved Ridge Estimators // Mathematics. 2024. Т. 12, № 19. № статьи 3027. DOI: 10.3390/math12193027.
3. *Alharthi M.F., Akhtar N.* Newly Improved Two-Parameter Ridge Estimators: A Better Approach for Mitigating Multicollinearity in Regression Analysis // Axioms. 2025. Т. 14, № 3. № статьи 186. DOI: 10.3390/axioms14030186.
4. *Ocaña-Riola R., Pérez-Romero C., Ortega-Díaz M.I. и др.* Multilevel Zero-One Inflated Beta Regression Model for the Analysis of the Relationship between Exogenous Health Variables and Technical Efficiency in the Spanish National Health System Hospitals // International Journal of Environmental Research and Public Health. 2021. Т. 18, № 19. № статьи 10166. DOI: 10.3390/ijerph181910166.
5. *Qiu M., Yang W., Wei F. et al.* A Topic Modeling Based on Prompt Learning // Electronics. 2024. Т. 13, № 16. № статьи 3212. DOI: 10.3390/electronics13163212.
6. *Bartol K., Bojanić D., Petković T. et al.* Linear Regression vs. Deep Learning: A Simple Yet Effective Baseline for Human Body Measurement // Sensors. 2022. Т. 22, № 5. № статьи 1885. DOI: 10.3390/s22051885.
7. *Kim M., Kim D.* A Suggestion on the LDA-Based Topic Modeling Technique Based on ElasticSearch for Indexing Academic Research Results // Applied Sciences. 2022. Т. 12, № 6. № статьи 3118. DOI: 10.3390/app12063118.
8. *Albrekht V., Mukhamediev R.I., Popova Y. et al.* Top2Vec Topic Modeling to Analyze the Dynamics of Publication Activity Related to Environmental Monitoring Using Unmanned Aerial Vehicles // Publications. 2025. Т. 13, № 2. № статьи 15. DOI: 10.3390/publications13020015.
9. *Родионов, Д.Г.* Тематическое моделирование информационной среды медиакомпаний: инструментальный комплекс LDA-TF-IDF / Д. Г. Родионов, Е. А. Конников, П. А. Пашинина, С. И. Шаныгин // Мягкие измерения и вычисления. – 2024. – Т. 76. – № 3. – С. 72-84. – DOI 10.36871/2618-9976.2024.03.006. – EDN COCJYG.
10. *Chen W., Rabhi F., Liao W. et al.* Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study // Electronics. 2023. Т. 12, № 12. № статьи 2605. DOI: 10.3390/electronics12122605.
11. *Родионов, Д. Г.* Квантификаторы информационной среды финансового рынка / Д. Г. Родионов, П. А. Пашинина, Е. А. Конников // Экономические науки. – 2022. – № 211. – С. 125-128. – DOI 10.14451/1.211.125. – EDN WBAPHW.
12. *Liu B.-H., Zhang L.-W., Wei Y.-Q. et al.* Dual Power Transformation and Yeo–Johnson Techniques for Static and Dynamic Reliability Assessments // Buildings. 2024. Т. 14, № 11. № статьи 3625. DOI: 10.3390/buildings14113625.
13. *Fang L., Zhou Z., Hong Y.* Symmetry Analysis of the Uncertain Alternative Box-Cox Regression Model // Symmetry. 2022. Т. 14, № 1. № статьи 22. DOI: 10.3390/sym14010022.
14. *Alshamrani S.S.* Machine Learning Techniques Improving the Box–Cox Transformation in Breast Cancer Prediction // Electronics. 2025. Т. 14, № 16. № статьи 3173. DOI: 10.3390/electronics14163173.
15. *Yuan Y., Yang K., Cheng L. et al.* Effect of Normalization Methods on Accuracy of Estimating Low- and High-Molecular Weight PAHs Distribution in the Soils of a Coking Plant // International Journal of Environmental Research and Public Health. 2022. Т. 19, № 23. № статьи 15470. DOI: 10.3390/ijerph192315470.
16. *Nayak S.K., Pradhan B., Mohanty B. et al.* Dimensionality Reduction Techniques for Heart Rate Variability Analysis [Электронный ресурс]. Encyclopedia. 2023. URL: https://encyclopedia.pub/entry/51982 (дата обращения: 10.12.2025).
17. *Конников, Е.А.* Совершенствование методов оценки устойчивости развития промышленных предприятий (октант устойчивости развития предприятия) / Е. А. Конников // Маркетинг менеджмент в цифровой экономике. – 2015. – Т. 1. – № 4. – С. 4-35. – EDN ZCGXSZ.
18. *Родионов, Д.Г.* Автоматизированный алгоритм системного анализа конкурентоспособности цифрового предприятия в рамках информационной среды / Д. Г. Родионов, Р. М. Мугутдинов, Е. А. Конников // Экономические науки. – 2021. – № 200. – С. 98-108. – DOI 10.14451/1.200.98. – EDN RRFFSY.
19. *Родионов, Д.Г.* Трансформация экологической среды социально-экономических систем под воздействием факторов информационной среды / Д. Г. Родионов, Е. А. Короткова, Д. А. Крыжко [и др.] // Экономические науки. – 2021. – № 201. – С. 98-111. – DOI 10.14451/1.201.98. – EDN GDIKCB.
20. *Marambakuyana W.A., Shongwe S.C.* Composite and Mixture Distributions for Heavy-Tailed Data—An Application to Insurance Claims // Mathematics. 2024. Т. 12, № 2. № статьи 335. DOI: 10.3390/math12020335.

METHOD OF INTEGRAL-QUANTILE REDUCTION OF THEMATIC PROPORTIONS OF THE LATENT DIRECTION MODEL TO NORMAL DISTRIBUTION WITH SUBSEQUENT SPHERICISATION BASED ON MIXTURES OF BETA DISTRIBUTIONS

© 2026 D.G. Rodionov, E.A. Konnikov, P.A. Polyakov

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

The distributions of topic shares obtained by the Latent Dirichlet Allocation (LDA) method usually deviate significantly from normal distributions. They are highly non-linear and U-shaped, concentrating around 0 and 1. This creates problems when using thematic features in linear and interpretable models, which assume symmetry and normality of data. This paper proposes a method for linearising LDA topic distributions based on probabilistic integral transformation using a mixture of beta distributions and subsequent probit transformation. This is followed by centring and spherical whitening transformation of the features. The proposed method significantly evens out the distributions of thematic features, bringing them closer to the normal form. As a result, in regression models of the form "target variable ~ topics" there is a 28% increase in the coefficient of determination R^2 relative to the original features and a decrease in the mean square error compared to models based on unprocessed features. In addition, the fit to the Gaussian-Markov assumptions improves. The heteroscedasticity of the residuals is reduced and the multicollinearity of the features is eliminated. The presented approach expands the possibilities of text processing. It increases the interpretability of thematic models and facilitates the inclusion of LDA topics in Bayesian and classical linear models for forecasting and analysis tasks.

Keywords: LDA, beta distribution, probabilistic integral transformation, probit, whitening, data normalization, linear regression, thematic modeling.

DOI: 10.37313/1990-5378-2026-28-2-203-210
EDN: HZWCGC

The work was carried out as part of the project "Development of a methodology for forming an instrumental base for analysing and modelling the spatial socio-economic development of systems in the context of digitalisation based on internal reserves" (FSEG-2023-0008).

REFERENCES

1. HandWiki. Heteroscedasticity [Elektronnyj resurs]. Encyclopedia. 2022. URL: https://encyclopedia.pub/entry/28997 (data obrashcheniya: 10.12.2025).
2. *Akhtar N., Alharthi M.F., Khan M.S.* Mitigating Multicollinearity in Regression: A Study on Improved Ridge Estimators // Mathematics. 2024. T. 12, № 19. № stat'i 3027. DOI: 10.3390/math12193027.
3. *Alharthi M.F., Akhtar N.* Newly Improved Two-Parameter Ridge Estimators: A Better Approach for Mitigating Multicollinearity in Regression Analysis // Axioms. 2025. T. 14, № 3. № stat'i 186. DOI: 10.3390/axioms14030186.
4. *Ocaña-Riola R., Pérez-Romero C., Ortega-Díaz M.I. et al.* Multilevel Zero-One Inflated Beta Regression Model for the Analysis of the Relationship between Exogenous Health Variables and Technical Efficiency in the Spanish National Health System Hospitals // International Journal of Environmental Research and Public Health. 2021. T. 18, № 19. № stat'i 10166. DOI: 10.3390/ijerph181910166.
5. *Qiu M., Yang W., Wei F. et al.* A Topic Modeling Based on Prompt Learning // Electronics. 2024. T. 13, № 16. № stat'i 3212. DOI: 10.3390/electronics13163212.
6. *Bartol K., Bojanić D., Petković T. et al.* Linear Regression vs. Deep Learning: A Simple Yet Effective Baseline for Human Body Measurement // Sensors. 2022. T. 22, № 5. № stat'i 1885. DOI: 10.3390/s22051885.
7. *Kim M., Kim D.* A Suggestion on the LDA-Based Topic Modeling Technique Based on ElasticSearch for Indexing Academic Research Results // Applied Sciences. 2022. T. 12, № 6. № stat'i 3118. DOI: 10.3390/app12063118.
8. *Albrekht V., Mukhamediev R.I., Popova Y. et al.* Top2Vec Topic Modeling to Analyze the Dynamics of Publication Activity Related to Environmental Monitoring Using Unmanned Aerial Vehicles // Publications. 2025. T. 13, № 2. № stat'i 15. DOI: 10.3390/publications13020015.
9. *Rodionov, D.G.* Tematicheskoe modelirovanie informacionnoj sredy mediakompanij: instrumental'nyj kompleks LDA-TF-IDF / D. G. Rodionov, E. A. Konnikov, P. A. Pashinina, S. I. Shanygin // Myagkie izmereniya i vychisleniya. – 2024. – T. 76. – № 3. – S. 72-84. – DOI 10.36871/2618-9976.2024.03.006. – EDN COCJYG.
10. *Chen W., Rabhi F., Liao W. et al.* Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study // Electronics. 2023. T. 12, № 12. № stat'i 2605. DOI: 10.3390/electronics12122605.
11. *Rodionov, D. G.* Kvantifikatory informacionnoj sredy finansovogo rynka / D. G. Rodionov, P. A. Pashinina, E. A. Konnikov // Ekonomicheskie nauki. – 2022. – № 211. – S. 125-128. – DOI 10.14451/1.211.125. – EDN WBAPHW.
12. *Liu B.-H., Zhang L.-W., Wei Y.-Q. et al.*Dual Power Transformation and Yeo–Johnson Techniques for Static and Dynamic Reliability Assessments // Buildings. 2024. T. 14, № 11. № stat'i 3625. DOI: 10.3390/buildings14113625.
13. *Fang L., Zhou Z., Hong Y.* Symmetry Analysis of the Uncertain Alternative Box-Cox Regression Model // Symmetry. 2022. T. 14, № 1. № stat'i 22. DOI: 10.3390/sym14010022.
14. *Alshamrani S.S.* Machine Learning Techniques Improving the Box–Cox Transformation in Breast Cancer Prediction // Electronics. 2025. T. 14, № 16. № stat'i 3173. DOI: 10.3390/electronics14163173.

15. *Yuan Y., Yang K., Cheng L. et al.* Effect of Normalization Methods on Accuracy of Estimating Low- and High-Molecular Weight PAHs Distribution in the Soils of a Coking Plant // International Journal of Environmental Research and Public Health. 2022. T. 19, № 23. № stat'i 15470. DOI: 10.3390/ijerph192315470.
16. *Nayak S.K., Pradhan B., Mohanty B. et al.* Dimensionality Reduction Techniques for Heart Rate Variability Analysis [Elektronnyj resurs]. Encyclopedia. 2023. URL: https://encyclopedia.pub/entry/51982 (data obrashcheniya: 10.12.2025).
17. *Konnikov, E.A.* Sovershenstvovanie metodov ocenki ustojchivosti razvitiya promyshlennyh predpriyatij (oktant ustojchivosti razvitiya predpriyatiya) / E. A. Konnikov // Marketing menedzhment v cifrovoj ekonomike. – 2015. – T. 1. – № 4. – S. 4-35. – EDN ZCGXSZ.
18. *Rodionov, D.G.* Avtomatizirovannyj algoritm sistemnogo analiza konkurentosposobnosti cifrovogo predpriyatiya v ramkah informacionnoj sredy / D. G. Rodionov, R. M. Mugutdinov, E. A. Konnikov // Ekonomicheskie nauki. – 2021. – № 200. – S. 98-108. – DOI 10.14451/1.200.98. – EDN RRFFSY.
19. *Rodionov, D.G.* Transformaciya ekologicheskoj sredy social'no-ekonomicheskih sistem pod vozdejstviem faktorov informacionnoj sredy / D. G. Rodionov, E. A. Korotkova, D. A. Kryzhko [i dr.] // Ekonomicheskie nauki. – 2021. – № 201. – S. 98-111. – DOI 10.14451/1.201.98. – EDN GDIKCB.
20. *Marambakuyana W.A., Shongwe S.C.* Composite and Mixture Distributions for Heavy-Tailed Data—An Application to Insurance Claims // Mathematics. 2024. T. 12, № 2. № stat'i 335. DOI: 10.3390/math12020335.

Dmitry Rodionov, Doctor of Economics, Director of the Higher School of Engineering and Economics.
E-mail: drodionov@spbstu.ru
Evgeny Konnikov, Ph.D. in Economics, Associate Professor at the Higher School of Engineering and Economics, Head of the Research Laboratory «Polytech-Invest,» and Academic Supervisor of the Master's Program 01.04.0503 "Neurostatistical Technologies in Marketing". E-mail: konnikov_ea@spbstu.ru
Prohor Polyakov, Laboratory Assistant at the Polytech-Invest Research Laboratory. E-mail: prohor@polyakov-box.ru