НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-999-1001
УДК 004.961

Вероятностный метод матричной кластеризации с априорным распределением признаков для формирования несмещенной контрольной группы

Дмитрий Андреевич Усольцев✉

Институт геномной медицины, Детская больница Нейшенвайд, Колумбус, 43205, США
Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

dusoltsev.27@gmail.com✉, https://orcid.org/0000-0001-8072-310X

Аннотация

Предложен метод вероятностной кластеризации матричных данных с использованием априорного распределения признаков и понижения размерности (Singular Value Decomposition, SVD). Метод позволяет выделять в большой контрольной группе кластер, статистически сопоставимый с тестовой группой, что снижает систематические искажения при дальнейшем сравнительном анализе. Показано, что предлагаемый метод позволяет корректно подбирать контрольную группу в случаях, когда известный метод ближайшего соседа дает ложноположительные результаты. Представленный метод применялся для отбора контрольных групп в исследованиях на основе медико-генетической базы данных, проводимых в Национальном медицинском исследовательском центре имени В.А. Алмазова.

Ключевые слова

матричная кластеризация, SVD, априорное распределение, расстояние Махаланобиса, χ^2-критерий

Ссылка для цитирования: Усольцев Д.А. Вероятностный метод матричной кластеризации с априорным распределением признаков для формирования несмещенной контрольной группы // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 999–1001. doi: 10.17586/2226-1494-2025-25-5-999-1001

Probabilistic matrix clustering with feature priors for unbiased control selection

Dmitrii A. Usoltsev✉

Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, 43205, USA
ITMO University, Saint Petersburg, 197101, Russian Federation

dusoltsev.27@gmail.com✉, https://orcid.org/0000-0001-8072-310X

Abstract

We propose a probabilistic matrix-clustering method that leverages a prior distribution of features and dimensionality reduction (Singular Value Decomposition, SVD). The approach identifies, within a large control pool, a cluster statistically comparable to the test cohort, thereby reducing systematic bias in downstream comparative analyses. We show that the method correctly selects control groups in scenarios where standard nearest-neighbor matching produces false positives. The method has been used to construct control groups in studies based on the Russian Biobank at the Almazov National Medical Research Centre (Ministry of Health of the Russian Federation).

Keywords

matrix clustering, SVD, prior (feature) distribution, Mahalanobis distance, χ^2-criterion

For citation: Usoltsev D.A. Probabilistic matrix clustering with feature priors for unbiased control selection. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 999–1001 (in Russian). doi: 10.17586/2226-1494-2025-25-5-999-1001

В исследованиях по типу «случай-контроль» требуется корректно сопоставить тестовую группу пациентов (с заболеванием/фенотипом) с контрольной группой большой размерности [1]. Несовпадение распределений ключевых факторов между группами приводит к систематическим ошибкам при сравнительном анализе [2]. Распространенные методы подбора контрольной группы — ближайшие соседи по евклидову расстоянию и по расстоянию Махаланобиса [3] — в условиях высокой размерности часто дают сбои: возвращают ограниченное число сопоставимых наблюдений (образцов) и часто подбирают нерелевантные (ложноположительные) наблюдения. В настоящей работе предлагается формализовать подбор контрольной группы как вероятностную задачу на матрицах признаков $\mathbf{A}_1 \in \mathbb{R}^{n_1 \times m}$ и $\mathbf{A}_2 \in \mathbb{R}^{n_2 \times m}$, где $\mathbf{A}_1$ — тестовая матрица; $\mathbf{A}_2$ — контрольная матрица; n_1 и n_2 — число наблюдений в матрицах $\mathbf{A}_1$ и $\mathbf{A}_2$ соответственно; m — число характеристик.

Пусть требуется найти подмножество $\hat{S}_2 \subseteq \mathbf{A}_2$ такое, чтобы проекции тестовой и выбранной контрольной групп не отличались по ключевым статистикам и корреляциям в общем пространстве признаков, тем самым минимизируя смещения при последующем сравнительном анализе.

На первом этапе задача сводится к приведению двух наборов данных — тестового и контрольного — к единому признаковому пространству. Для этого проводится стандартизация признаков. В тестовом наборе $\mathbf{A}_1 \in \mathbb{R}^{n_1 \times m}$ вычисляются средние значения по каждому столбцу. Эти значения затем вычитаются из обеих матриц $\mathbf{A}_1$ и $\mathbf{A}_2 \in \mathbb{R}^{n_2 \times m}$, что обеспечивает выравнивание выборок по среднему значению признаков и устраняет глобальные линейные смещения.

Далее применяется такой метод снижения размерности, как сингулярное разложение матриц [4], позволяющий выделить наиболее информативные направления изменения данных. К матрице $\mathbf{A}_1$ применяется сингулярное разложение: $\mathbf{A}_1 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, где $\mathbf{U} \in \mathbb{R}^{n_1 \times n_1}$ — матрица левых сингулярных векторов; $\mathbf{\Sigma}$ — диагональная матрица сингулярных чисел; $\mathbf{V} \in \mathbb{R}^{m \times m}$ — матрица правых сингулярных векторов. В матрице $\mathbf{U}$ выбираются первые k столбцов, соответствующие наибольшим сингулярным числам. Полученная матрица $\hat{\mathbf{U}} \in \mathbb{R}^{n_1 \times k}$ определяет проекцию в подпространство наибольшей дисперсии. Проекции обоих наборов данных вычисляются следующим образом: $\hat{\mathbf{A}}_1 = \hat{\mathbf{U}}\mathbf{A}_1$ и $\hat{\mathbf{A}}_2 = \hat{\mathbf{U}}\mathbf{A}_2$. Каждые строки матриц $\hat{\mathbf{A}}_1$ и $\hat{\mathbf{A}}_2$ теперь представляют наблюдение в новом пространстве признаков размерности k общем для обеих выборок. Это обеспечивает сопоставимость полученных наборов данных.

После проекции на пространство главных компонент $\hat{\mathbf{A}}_1$, компоненты тестовой выборки рассматриваются как реализация случайной величины, подчиняющейся многомерному нормальному распределению. Оцениваются вектор математических ожиданий $\mathbf{m} \in \mathbb{R}^k$ и ковариационная матрица $\mathbf{\Sigma} \in \mathbb{R}^{k \times k}$, задающие априорное распределение признаков: $\hat{\mathbf{A}}_1 \sim \mathcal{N}(m, \mathbf{\Sigma})$.

Далее формулируется вероятностный критерий принадлежности. Для каждого объекта $x \in \hat{\mathbf{A}}_2$ вычисляется расстояние Махаланобиса как мера правдоподобия его принадлежности к распределению тестовой выборки: $D_M^2(x) = (x - m)^T\mathbf{\Sigma}^{-1}(x - m)$. Таким образом, каждый объект из контрольной выборки получает количественную оценку степени принадлежности к априорному распределению. В отличие от детерминированного отнесения объектов к кластерам, используется вероятностный критерий: объект включается в кластер $\hat{S}_2$, если значение $D_M(x)$ не превышает порог, соответствующий 95%-квантилю распределения χ^2 с k степенями свободы. Это соответствует включению в доверительную область уровня 0,95 для многомерного нормального распределения. Таким образом, кластер $\hat{S}_2 \subseteq \mathbf{A}_2$ формируется как $\hat{S}_2 = \{x \in \mathbf{A}_2 | D_M(\hat{\mathbf{U}}^T x) \le \chi^2_{k,1-\alpha}\}$, где α — уровень значимости. Неравенство означает, что x принадлежит 95%-ной доверительной области (при $\alpha = 0{,}05$) распределения $\mathcal{N}(m, \mathbf{\Sigma})$, которая в k-мерном пространстве имеет форму эллипсоида (в двумерном случае — эллипса).

Для воспроизводимой оценки предложенного метода было проведено два эксперимента. В эксперименте 1 тестовая и контрольная выборки были симулированы из одного многомерного нормального распределения, в эксперименте 2 — из разных. Предполагается, что практически все образцы из контрольной выборки будут отобраны предлагаемым методом в эксперименте 1, и никакие из образцов не будут отобраны во эксперименте 2. Образцы, отобранные в эксперименте 2, считаются ложноположительными, так как относятся к другому распределению.

В эксперименте 1 была промоделирована тестовая матрица $\mathbf{A}_1 \in \mathbb{R}^{n_1 \times m}$, где $n_1 = 200$ и $m = 30$, как выборка из многомерного нормального распределения $\mathcal{N}(0, \mathbf{\Sigma})$. В качестве матрицы ковариаций $\mathbf{\Sigma}$ с параметром корреляции Пирсона r [5] была использована матрица Тёплица [6] размером $m \times m$, где каждый элемент равен $\mathbf{\Sigma}_{ij} = r^{|i-j|}$. Такая структура ковариаций была выбрана для того, чтобы промоделировать корреляции, характерные для реальных данных. Контрольная матрица $\mathbf{A}_2 \in \mathbb{R}^{n_2 \times m}$, $n_2 = 20\,000$ генерировалась из того же распределения, что и матрица $\mathbf{A}_1$. В эксперименте 2 тестовый набор был тот же, а контрольная матрица смещалась на константу равную 2,7 относительно тестового набора в пространстве главных компонент, которая была выбрана экспериментально. Затем в обоих экспериментах был применен предлагаемый метод с целью поиска в матрице $\mathbf{A}_2$ контрольной подгруппы. Декомпозиция Singular Value Decomposition (SVD) была применена к матрице $\mathbf{A}_1$ и обе матрицы проектировались в k-мерное подпространство ($k = 9$), объясняющее более 70 % дисперсии. В этом пространстве для тестовой группы оценивались вектор средних $\boldsymbol{\mu}$ и ковариационная матрица $\mathbf{\Sigma}$, и предлагаемый метод отбирал все наблюдения $x \in \mathbf{A}_2$, удовлетворяющие вероятностному критерию $D_M^2(x) = (x - m)^T\mathbf{\Sigma}^{-1}(x - m) \le \chi^2_{k,1-\alpha}$, где $\alpha = 0{,}05$.

В качестве метода для сравнения надежности отбора контрольной группы был использован метод ближайшего соседа по расстоянию Махаланобиса в исходном пространстве признаков. По объединенным данным $[\mathbf{A}_1; \mathbf{A}_2]$ строилась матрица попарных расстояний Махаланобиса, после чего выполнялось сопоставление жадным алгоритмом каждого тестового наблюдения с

ближайшим доступным контрольным наблюдением по формуле «один к одному».

Предлагаемый метод показал более корректные результаты по сравнению с методом ближайшего соседа. В эксперименте 1 предлагаемый метод сформировал широкий набор (N = 19 744) статистически сопоставимых контрольных наблюдений, тогда как метод ближайшего соседа вернул число контрольных наблюдений равное числу тестовых наблюдений (N = 200). В эксперименте 2 предлагаемый метод не включил ни одного наблюдения из контрольной группы, избегая ложноположительных совпадений. Напротив, метод ближайшего соседа ошибочно отобрал N = 200 контрольных наблюдений, формируя ложные пары с тестовой выборкой.

В результате следует, что в настоящей работе был представлен метод формирования несмещенной контрольной выборки для исследований по типу «случай-контроль» [1]. Комбинация SVD и вероятностного отбора по Махаланобису с χ^2-критерием обеспечивает статистическую сопоставимость групп и улучшает корректность выводов. Предлагаемый метод внедрен в исследования, проводимые в Национальном медицинском исследовательском центре имени В.А. Алмазова. В частности, он позволил отобрать контрольную группу из медико-генетической базы данных для сравнительного исследования межпоколенческих эффектов потомков жителей блокадного Ленинграда [7].

Литература

1. Artomov M., Loboda A.A., Artyomov M.N., Daly M.J. Public platform with 39,472 exome control samples enables association studies without genotype sharing // Nature Genetics. 2024. V. 56. N 2. P. 327–335. https://doi.org/10.1038/s41588-023-01637-y
2. Pearce N. Analysis of matched case-control studies // BMJ Online. 2016. V. 352. P. i969. https://doi.org/10.1136/bmj.i969
3. Ghosh A., Ghosh A.K., SahaRay R., Sarkar S. Classification using global and local Mahalanobis distances // Journal of Multivariate Analysis. 2025. V. 207. P. 105417. https://doi.org/10.1016/j.jmva.2025.105417
4. Brunton S.L., Kutz J.N. Singular Value Decomposition (SVD) // Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. 2019. P. 3–46. https://doi.org/10.1017/9781108380690.002
5. Rovetta A. Raiders of the lost correlation: a guide on using pearson and spearman coefficients to detect hidden correlations in medical sciences // Cureus. 2020. V. 12. N 11. P. e11794. https://doi.org/10.7759/cureus.11794
6. Wang Z., Li G., Hu F., Chi N. Toeplitz concatenated matrix aided ICA algorithm for super-Nyquist multiband CAP VLC systems // Optics Express. 2020. V. 28. N 20. P. 29876–29894. https://doi.org/10.1364/OE.404925
7. Tolkunova K., Usoltsev D., Moguchaia E., Boyarinova M., Kolesova E., Erina A., et al. Transgenerational and intergenerational effects of early childhood famine exposure in the cohort of offspring of Leningrad Siege survivors // Scientific Reports. 2023. V. 13. N 1. P. 11188. https://doi.org/10.1038/s41598-023-37119-8

References

1. Artomov M., Loboda A.A., Artyomov M.N., Daly M.J. Public platform with 39,472 exome control samples enables association studies without genotype sharing. *Nature Genetics*, 2024, vol. 56, no. 2, pp. 327–335. https://doi.org/10.1038/s41588-023-01637-y
2. Pearce N. Analysis of matched case-control studies. *BMJ Online*, 2016, vol. 352, pp. i969. https://doi.org/10.1136/bmj.i969
3. Ghosh A., Ghosh A.K., SahaRay R., Sarkar S. Classification using global and local Mahalanobis distances. *Journal of Multivariate Analysis*, 2025, vol. 207, pp. 105417. https://doi.org/10.1016/j.jmva.2025.105417
4. Brunton S.L., Kutz J.N. Singular Value Decomposition (SVD). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*, 2019, pp. 3–46. https://doi.org/10.1017/9781108380690.002
5. Rovetta A. Raiders of the lost correlation: a guide on using pearson and spearman coefficients to detect hidden correlations in medical sciences. *Cureus*, 2020, vol. 12, no. 11, pp. e11794. https://doi.org/10.7759/cureus.11794
6. Wang Z., Li G., Hu F., Chi N. Toeplitz concatenated matrix aided ICA algorithm for super-Nyquist multiband CAP VLC systems. *Optics Express*, 2020, vol. 28, no. 20, pp. 29876–29894. https://doi.org/10.1364/OE.404925
7. Tolkunova K., Usoltsev D., Moguchaia E., Boyarinova M., Kolesova E., Erina A., et al. Transgenerational and intergenerational effects of early childhood famine exposure in the cohort of offspring of Leningrad Siege survivors. *Scientific Reports*, 2023, vol. 13, no. 1, pp. 11188. https://doi.org/10.1038/s41598-023-37119-8

Автор

Усольцев Дмитрий Андреевич — старший научный сотрудник, Институт геномной медицины, Детская больница Нейшенвайд, Колумбус, 43205, США; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57279360300, https://orcid.org/0000-0001-8072-310X, dusoltsev.27@gmail.com

Author

Dmitrii A. Usoltsev — Senior Researcher, Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, 43205, USA; PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57279360300, https://orcid.org/0000-0001-8072-310X, dusoltsev.27@gmail.com

Статья поступила в редакцию 15.08.2025
Одобрена после рецензирования 05.09.2025
Принята к печати 21.09.2025

Received 15.08.2025
Approved after reviewing 05.09.2025
Accepted 21.09.2025