Methods for quantifying similarity of datasets

A Review, Taxonomy and Comparison

Marieke Stolte
Franziska Kappenberg, Jörg Rahnenführer, Andrea Bommert

TU Dortmund University
Department of Statistics

May 15, 2024

Motivation

Motivation and application in simulation studies

Quantifying the similarity between two or more datasets has widespread applications in statistics and machine learning:

- Generalizability of statistical models depends on similarity between datasets used for fitting and new datasets
- Meta-learning / transfer learning uses similarity to transfer insights for learning tasks between different datasets
- Two- or k-sample tests check whether the underlying distributions of two or more datasets coincide
- Similarity between simulated datasets and real datasets is crucial in simulation studies

Approach

- Extremely many approaches proposed in literature

Approach

- Extremely many approaches proposed in literature
- Goal: Review and comparison of more than 100 methods divided into 10 classes to guide choice of suitable method

Approach

- Extremely many approaches proposed in literature
- Goal: Review and comparison of more than 100 methods divided into 10 classes to guide choice of suitable method
- Criteria for inclusion of methods
  1. Method is applicable to multivariate data
  2. Method does not require any specific parametric or distributional assumptions (e.g. normal assumption)
  3. Method does not focus on a particular property of the data (e.g. means), but on the entire dataset or its entire distribution

Literature Review: Classes

Overview

- Comparison of cumulative distribution functions, density functions or characteristic functions
- Methods based on multivariate ranks
- Discrepancy measures for distributions
- Comparison based on summary statistics
- Different testing approaches
- Graph-based methods
- Methods based on inter-point distances
- Kernel-based methods
- Methods based on binary classification
- Distance and similarity measures for datasets

Overview

- Comparison of cumulative distribution functions, density functions or characteristic functions
- Methods based on multivariate ranks
- Discrepancy measures for distributions
- Comparison based on summary statistics
- Different testing approaches
- Graph-based methods
- Methods based on inter-point distances
- Kernel-based methods
- Methods based on binary classification
- Distance and similarity measures for datasets

Overview

- Comparison of cumulative distribution functions, density functions or characteristic functions
- Methods based on multivariate ranks
- Discrepancy measures for distributions
- Comparison based on summary statistics
- Different testing approaches
- Graph-based methods
- Methods based on inter-point distances
- Kernel-based methods
- Methods based on binary classification
- Distance and similarity measures for datasets

Overview

- Comparison of cumulative distribution functions, density functions or characteristic functions
- Methods based on multivariate ranks
- Discrepancy measures for distributions
- Comparison based on summary statistics
- Different testing approaches
- Graph-based methods
- Methods based on inter-point distances
- Kernel-based methods
- Methods based on binary classification
- Distance and similarity measures for datasets

Overview

- Comparison of cumulative distribution functions, density functions or characteristic functions
- Methods based on multivariate ranks
- Discrepancy measures for distributions
- Comparison based on summary statistics
- Different testing approaches
- Graph-based methods
- Methods based on inter-point distances
- Kernel-based methods
- Methods based on binary classification
- Distance and similarity measures for datasets

Discrepancy measures for distributions

- Distinction between probability metrics (fulfill all metric properties) and divergences (fulfill some metric properties)

Discrepancy measures for distributions

- Distinction between probability metrics (fulfill all metric properties) and divergences (fulfill some metric properties)
- Examples:

Discrepancy measures for distributions

- Distinction between probability metrics (fulfill all metric properties) and divergences (fulfill some metric properties)
- Examples:
  - Integral probability metrics (IPM, also called probability metrics with a ξ-structure):
    If distributions F_1, F_2 are identical, any function f has same expectation under both [20], so

$$IPM_{\mathcal{F}}(F_1, F_2) = \sup_{f \in \mathcal{F}} \left| \int f \, \mathrm{d}F_1 - \int f \, \mathrm{d}F_2 \right|,$$

    where $\mathcal{F}$ is a given set of functions.

Discrepancy measures for distributions

- Distinction between probability metrics (fulfill all metric properties) and divergences (fulfill some metric properties)
- Examples:
  - Integral probability metrics (IPM, also called probability metrics with a ξ-structure):
    If distributions F_1, F_2 are identical, any function f has same expectation under both [20], so

    $$IPM_{\mathcal{F}}(F_1, F_2) = \sup_{f \in \mathcal{F}} \left| \int f \, \mathrm{d}F_1 - \int f \, \mathrm{d}F_2 \right|,$$

    where $\mathcal{F}$ is a given set of functions.
  - f-divergences (also called Ali-Silvey distances or Csisár´s Φ-divergences):
    Identical distributions assign the same likelihood to every point [15], so

    $$D_f(F_1, F_2) = \int f\left(\frac{f_1(X)}{f_2(X)}\right) \mathrm{d}F_1,$$

    where $f : \mathbb{R}_+ \to \mathbb{R}$ convex continuous function such that $f(1) = 0$.
    E.g. Kullback-Leibler divergence [14] for $f = \log$.

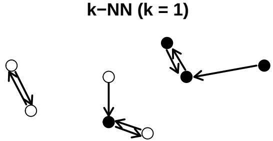
Graph-based methods

- Construct certain graph on the pooled sample

Graph-based methods

- Construct certain graph on the pooled sample
- Examples:

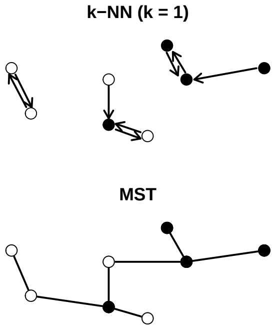
Graph-based methods



- Construct certain graph on the pooled sample
- Examples:
  - k-nearest neighbor (k-NN) graphs [7, 11, 12, 22]

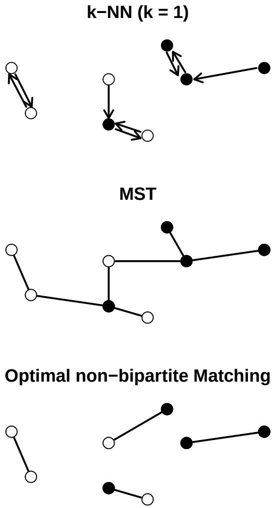
Graph-based methods

- Construct certain graph on the pooled sample
- Examples:
  - *k*-nearest neighbor (*k*-NN) graphs [7, 11, 12, 22]
  - Minimum spanning tree (MST) [6]



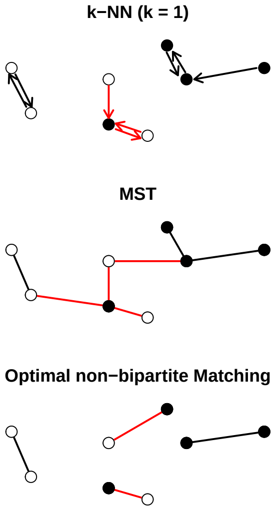
Graph-based methods

- Construct certain graph on the pooled sample
- Examples:
  - k-nearest neighbor (k-NN) graphs [7, 11, 12, 22]
  - Minimum spanning tree (MST) [6]
  - Optimal non-bipartite matching (cross-match test) [21]



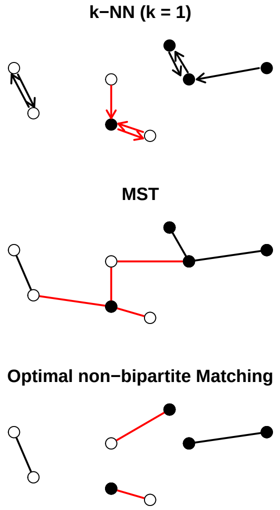
Graph-based methods

- Construct certain graph on the pooled sample
- Examples:
  - k-nearest neighbor (k-NN) graphs [7, 11, 12, 22]
  - Minimum spanning tree (MST) [6]
  - Optimal non-bipartite matching (cross-match test) [21]
- Count the edges that connect points from different datasets and use this edge count statistic or a normalized version of it



Graph-based methods

- Construct certain graph on the pooled sample
- Examples:
  - k-nearest neighbor (k-NN) graphs [7, 11, 12, 22]
  - Minimum spanning tree (MST) [6]
  - Optimal non-bipartite matching (cross-match test) [21]
- Count the edges that connect points from different datasets and use this edge count statistic or a normalized version of it
- If the datasets are similar, a high number of edges connecting points from different datasets is expected



Methods based on inter-point distances

- Theoretical justification [17]: the following two statements are equivalent
  - distributions of the samples ($\{X_i\}$ and $\{Y_i\}$) are equal
  - distributions of in-sample comparisons ($||X_i - X_j||$ and $||Y_i - Y_j||$) and distribution of between-sample comparisons ($||X_i - Y_j||$) are equal

$\Rightarrow$ Compare these distributions of in- and between-sample comparisons

Methods based on inter-point distances

- Theoretical justification [17]: the following two statements are equivalent
  - distributions of the samples ($\{X_i\}$ and $\{Y_i\}$) are equal
  - distributions of in-sample comparisons ($||X_i - X_j||$ and $||Y_i - Y_j||$) and distribution of between-sample comparisons ($||X_i - Y_j||$) are equal

  $\Rightarrow$ Compare these distributions of in- and between-sample comparisons
- Example: Energy statistic [1, 2, 26, 27] compares $2\times$ the mean of the between-sample distances to the sum of the means of the in-sample distances for both datasets

$$\mathcal{E}(X, Y) = 2\mathbb{E}(||X - Y||) - \mathbb{E}(||X - X'||) - \mathbb{E}(||Y - Y'||),$$

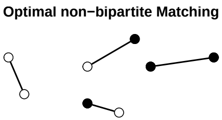
where $X, X' \overset{iid}{\sim} F_1$ and $Y, Y' \overset{iid}{\sim} F_2$

Method comparison

Criteria for method comparison

Applicability:

Cross-match test

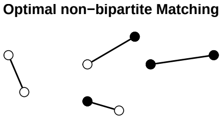


Criteria for method comparison

Applicability:

- Sensible inclusion of target variable?

Cross-match test



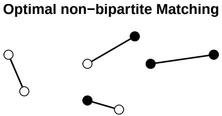
Criteria for method comparison

Applicability:

- Sensible inclusion of target variable?

Cross-match test

- No



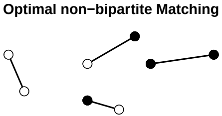
Criteria for method comparison

Applicability:

- Sensible inclusion of target variable?
- Numeric variables?

Cross-match test

- No



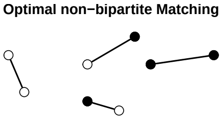
Criteria for method comparison

Applicability:

- Sensible inclusion of target variable?
- Numeric variables?

Cross-match test

- No
- Yes



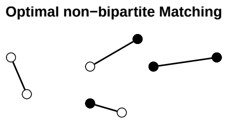
Criteria for method comparison

Applicability:

- Sensible inclusion of target variable?
- Numeric variables?
- Categorical variables?

Cross-match test

- No
- Yes



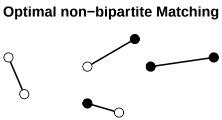
Criteria for method comparison

Applicability:

- Sensible inclusion of target variable?
- Numeric variables?
- Categorical variables?

Cross-match test

- No
- Yes
- No



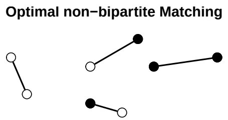
Criteria for method comparison

Applicability:

- Sensible inclusion of target variable?
- Numeric variables?
- Categorical variables?
- Unequal sample sizes permitted?

Cross-match test

- No
- Yes
- No



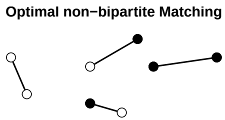
Criteria for method comparison

Applicability:

- Sensible inclusion of target variable?
- Numeric variables?
- Categorical variables?
- Unequal sample sizes permitted?

Cross-match test

- No
- Yes
- No
- Yes



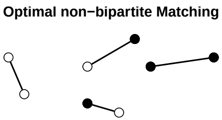
Criteria for method comparison

Applicability:

- Sensible inclusion of target variable?
- Numeric variables?
- Categorical variables?
- Unequal sample sizes permitted?
- $p > N$ permitted?

Cross-match test

- No
- Yes
- No
- Yes



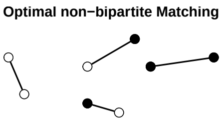
Criteria for method comparison

Applicability:

- Sensible inclusion of target variable?
- Numeric variables?
- Categorical variables?
- Unequal sample sizes permitted?
- $p > N$ permitted?

Cross-match test

- No
- Yes
- No
- Yes
- Yes



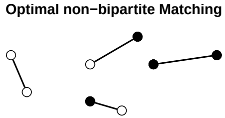
Criteria for method comparison

Applicability:

- Sensible inclusion of target variable?
- Numeric variables?
- Categorical variables?
- Unequal sample sizes permitted?
- $p > N$ permitted?
- Applicable to more than two datasets at a time ($k > 2$)?

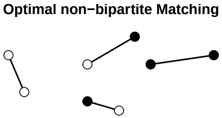
Cross-match test

- No
- Yes
- No
- Yes
- Yes



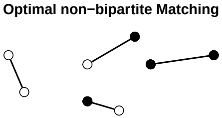
Criteria for method comparison

Applicability:	Cross-match test
Sensible inclusion of target variable?	No
Numeric variables?	Yes
Categorical variables?	No
Unequal sample sizes permitted?	Yes
$p > N$ permitted?	Yes
Applicable to more than two datasets at a time ($k > 2$)?	No



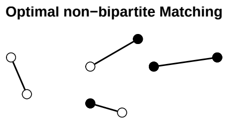
Criteria for method comparison

Applicability:	Cross-match test
Sensible inclusion of target variable?	No
Numeric variables?	Yes
Categorical variables?	No
Unequal sample sizes permitted?	Yes
$p > N$ permitted?	Yes
Applicable to more than two datasets at a time ($k > 2$)?	No
No additional training data / train test split required?	



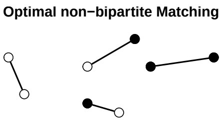
Criteria for method comparison

Applicability:	Cross-match test
Sensible inclusion of target variable?	No
Numeric variables?	Yes
Categorical variables?	No
Unequal sample sizes permitted?	Yes
$p > N$ permitted?	Yes
Applicable to more than two datasets at a time ($k > 2$)?	No
No additional training data / train test split required?	Yes



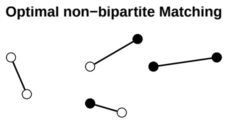
Criteria for method comparison

Applicability:	Cross-match test
Sensible inclusion of target variable?	No
Numeric variables?	Yes
Categorical variables?	No
Unequal sample sizes permitted?	Yes
$p > N$ permitted?	Yes
Applicable to more than two datasets at a time ($k > 2$)?	No
No additional training data / train test split required?	Yes
No further assumptions on distributions required?	



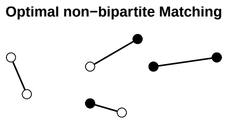
Criteria for method comparison

Applicability:	Cross-match test
Sensible inclusion of target variable?	No
Numeric variables?	Yes
Categorical variables?	No
Unequal sample sizes permitted?	Yes
$p > N$ permitted?	Yes
Applicable to more than two datasets at a time ($k > 2$)?	No
No additional training data / train test split required?	Yes
No further assumptions on distributions required?	No



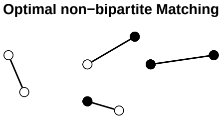
Criteria for method comparison

Applicability:	Cross-match test
Sensible inclusion of target variable?	No
Numeric variables?	Yes
Categorical variables?	No
Unequal sample sizes permitted?	Yes
$p > N$ permitted?	Yes
Applicable to more than two datasets at a time ($k > 2$)?	No
No additional training data / train test split required?	Yes
No further assumptions on distributions required?	No
No tuning / choice of additional parameters required?	



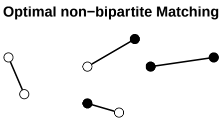
Criteria for method comparison

Applicability:	Cross-match test
Sensible inclusion of target variable?	No
Numeric variables?	Yes
Categorical variables?	No
Unequal sample sizes permitted?	Yes
$p > N$ permitted?	Yes
Applicable to more than two datasets at a time ($k > 2$)?	No
No additional training data / train test split required?	Yes
No further assumptions on distributions required?	No
No tuning / choice of additional parameters required?	Yes



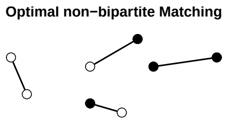
Criteria for method comparison

Applicability:	Cross-match test
Sensible inclusion of target variable?	No
Numeric variables?	Yes
Categorical variables?	No
Unequal sample sizes permitted?	Yes
$p > N$ permitted?	Yes
Applicable to more than two datasets at a time ($k > 2$)?	No
No additional training data / train test split required?	Yes
No further assumptions on distributions required?	No
No tuning / choice of additional parameters required?	Yes
Implemented in any software?	



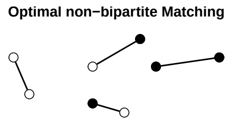
Criteria for method comparison

Applicability:	Cross-match test
Sensible inclusion of target variable?	No
Numeric variables?	Yes
Categorical variables?	No
Unequal sample sizes permitted?	Yes
$p > N$ permitted?	Yes
Applicable to more than two datasets at a time ($k > 2$)?	No
No additional training data / train test split required?	Yes
No further assumptions on distributions required?	No
No tuning / choice of additional parameters required?	Yes
Implemented in any software?	Yes



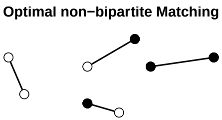
Criteria for method comparison

Applicability:	Cross-match test
Sensible inclusion of target variable?	No
Numeric variables?	Yes
Categorical variables?	No
Unequal sample sizes permitted?	Yes
$p > N$ permitted?	Yes
Applicable to more than two datasets at a time ($k > 2$)?	No
No additional training data / train test split required?	Yes
No further assumptions on distributions required?	No
No tuning / choice of additional parameters required?	Yes
Implemented in any software?	Yes
Computational complexity?	



Criteria for method comparison

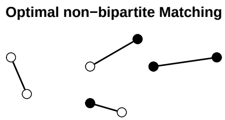
Applicability:	Cross-match test
Sensible inclusion of target variable?	No
Numeric variables?	Yes
Categorical variables?	No
Unequal sample sizes permitted?	Yes
$p > N$ permitted?	Yes
Applicable to more than two datasets at a time ($k > 2$)?	No
No additional training data / train test split required?	Yes
No further assumptions on distributions required?	No
No tuning / choice of additional parameters required?	Yes
Implemented in any software?	Yes
Computational complexity?	$\mathcal{O}(N^3)$



Criteria for method comparison

Interpretability:

Cross-match test

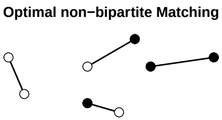


Criteria for method comparison

Interpretability:

- Interpretable units?

Cross-match test



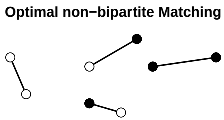
Criteria for method comparison

Interpretability:

- Interpretable units?

Cross-match test

- Yes



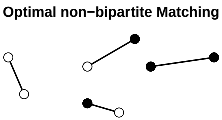
Criteria for method comparison

Interpretability:

- Interpretable units?
- Lower bound?
- Upper bound?

Cross-match test

- Yes



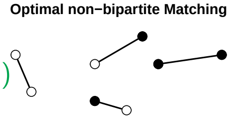
Criteria for method comparison

Interpretability:

- Interpretable units?
- Lower bound?
- Upper bound?

Cross-match test

- Yes
- 0
- $\min(n_1, n_2)$



Criteria for method comparison

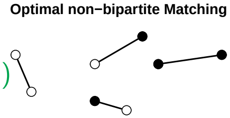
Interpretability:

- Interpretable units?
- Lower bound?
- Upper bound?

Theoretical properties:

Cross-match test

- Yes
- 0
- $\min(n_1, n_2)$



Criteria for method comparison

Interpretability:

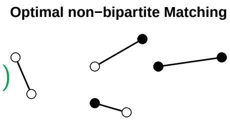
- Interpretable units?
- Lower bound?
- Upper bound?

Theoretical properties:

- Rotation invariant?
- Location change invariant?
- Scale invariant?

Cross-match test

- Yes
- 0
- $\min(n_1, n_2)$



Criteria for method comparison

Interpretability:

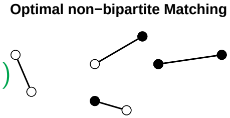
- Interpretable units?
- Lower bound?
- Upper bound?

Theoretical properties:

- Rotation invariant?
- Location change invariant?
- Scale invariant?

Cross-match test

- Yes
- 0
- $\min(n_1, n_2)$



- Yes
- Yes
- Yes

Criteria for method comparison

Interpretability:

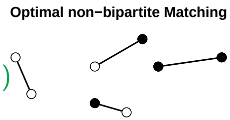
- Interpretable units?
- Lower bound?
- Upper bound?

Theoretical properties:

- Rotation invariant?
- Location change invariant?
- Scale invariant?
- Positive definite?
- Symmetric?
- Triangle inequality?

Cross-match test

- Yes
- 0
- $\min(n_1, n_2)$



- Yes
- Yes
- Yes

Criteria for method comparison

Interpretability:

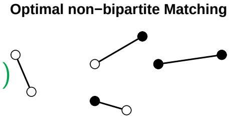
- Interpretable units?
- Lower bound?
- Upper bound?

Theoretical properties:

- Rotation invariant?
- Location change invariant?
- Scale invariant?
- Positive definite?
- Symmetric?
- Triangle inequality?

Cross-match test

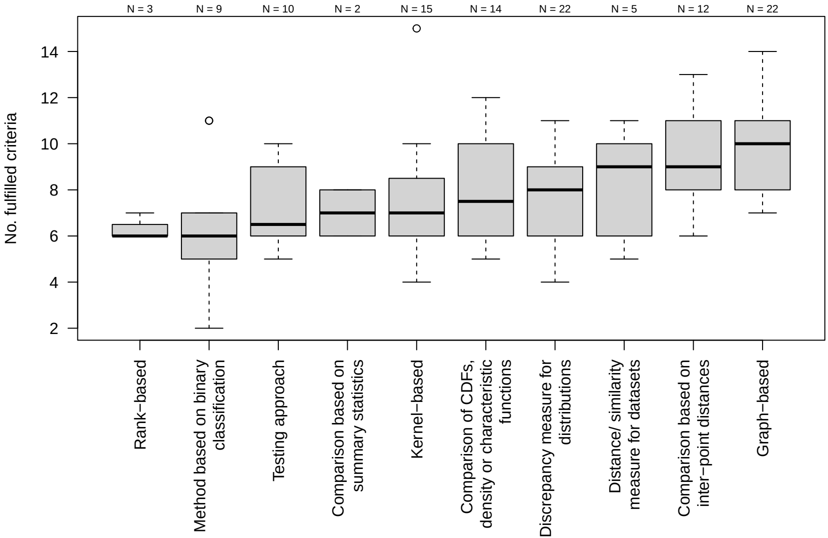
- Yes
- 0
- $\min(n_1, n_2)$



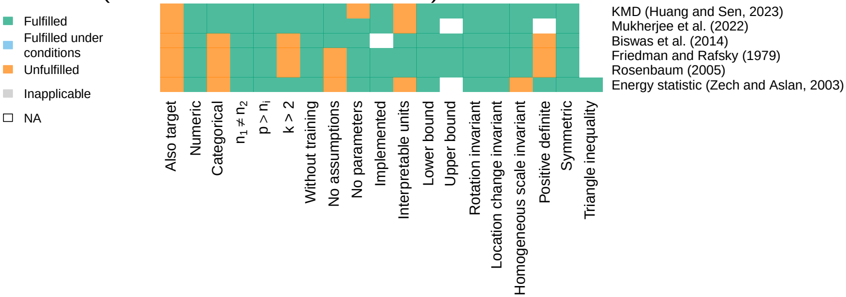
- Yes
- Yes
- Yes
- No
- Yes
- Unknown

Results

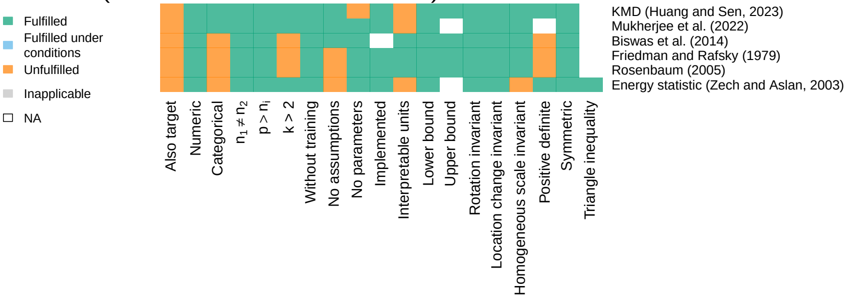
Number of fulfilled criteria



Top 6 methods ($\geq$ 13 criteria fulfilled)



Top 6 methods ($\geq$ 13 criteria fulfilled)



1. KMD: kernel-based test using the association between the features and the sample membership to quantify the dissimilarity of multiple distributions [13]
2. Mukherjee et al. (2022): graph-based test using non-bipartite optimal matchings [19]
3. Biswas et al. (2014): graph-based test using the shortest Hamiltonian path [3]
4. Friedman and Rafsky (1979): Friedman-Rafsky test, based on minimal spanning tree [6]
5. Rosenbaum (2005): cross-match test, based on non-bipartite optimal matchings [21]
6. Zech and Aslan (2003): Energy statistic [27]

Interactive Online Result Table

Depending on application some of the criteria are mandatory, others negligible

$\Rightarrow$ online tool which allows custom filtering and sorting

Comparison of Methods for Quantifying Dataset Similarity

The following interactive table contains the results of the article "Comparison of Methods for Quantifying Dataset Similarity" (DOI: ...). In this, 114 methods were compared with respect to the criteria specified in the columns here. For details on the comparisons and additional information, please refer to the article. If you have additions or corrections to the criteria for a method, please add an issue in the corresponding [GitHub repository]. We will be happy to incorporate them here.

Legend

✓ Fulfilled (✓) Fulfilled for certain parameter choices ✗ Unfulfilled -- Inapplicable

Column visibility ▾

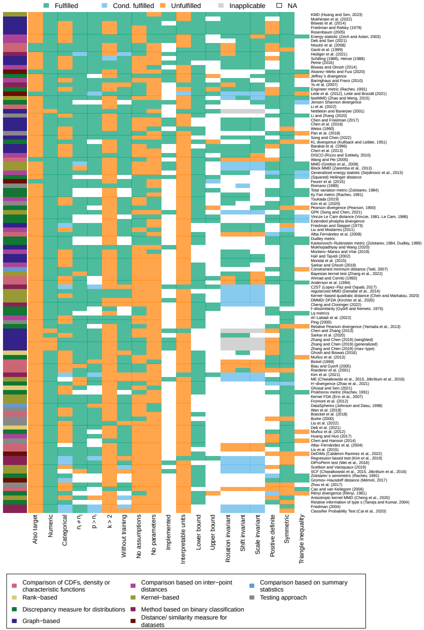
Search:

Method/ Article	Also target?	Numeric?	Categorical?	$n_i \neq n_j$?	$p>n_i$?	$k>2$?	Without training?	No assumptions?	No parameters?	Implemented?	Complexity?	Interpretable units?	Lower bound?	Upper bound?	Rotation invariant?	Location change invariant?	Homogeneous scale invariant?	Positive definite?	Symmetric?	Triangle inequality?	No. fulfilled criteria	No. cond. fulfilled criteria	No. unfulfilled criteria	No. NAs	Class	Subclass	Link
All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All
Energy statistic (Zech and Aslan, 2003)	✗	✓	✗	✓	✓	✓	✓	✗	✓	✓	$O(N^2)$	✗	✓	NA	✓	✓	✗	✓	✓	✓	13	0	5	1	Comparison based on ...	Comparison based on ...	Energy statistic (Zech and Aslan, 2003)
Generalized energy statistic (Sejdinovic et al., 2013)	✗	✓	✓	✓	✓	✗	✓	✗	✗	NA	NA	✗	✓	NA	✓	✓	NA	(✓)	✓	(✓)	9	2	5	3	Comparison based on ...	Comparison based on ...	Generalized energy statistic (Sejdinovic et al., 2013)
DISCO (Rizzo and Székely, 2010)	✗	✓	✗	✓	✓	✓	✓	✗	✗	✓	$O(N^2)$	✗	✓	NA	✓	✓	✗	NA	✓	NA	10	0	6	3	Comparison based on ...	Comparison based on ...	DISCO (Rizzo and Székely, 2010)
Huang and Huo (2017)	✗	✓	✗	✓	✓	✗	✓	✗	✗	NA	$O(mN\log N)$	✗	✓	NA	NA	NA	NA	✗	✓	NA	6	0	7	6	Comparison based on ...	Comparison based on ...	Huang and Huo (2017)
Deb and Sen (2021)	✗	✓	✗	✓	✓	✓	✓	✗	✓	✓	$O(N^3)+O(n_1n_2p)$	✗	✓	NA	NA	✓	✓	✓	✓	NA	12	0	4	3	Comparison based on ...	Comparison based on ...	Deb and Sen (2021)
Al-Labadi et al. (2022)	✗	✓	✗	✓	✓	✓	✓	✗	✗	NA	NA	✗	✓	NA	NA	NA	NA	NA	✓	NA	7	0	5	7	Comparison based on ...	Comparison based on ...	Al-Labadi et al. (2022)
Baringhaus and Franz (2010)	✗	✓	✗	✓	✓	✓	✓	✗	✗	✓	NA	✗	✓	NA	✓	✓	NA	✓	✓	NA	11	0	5	3	Comparison based on ...	Comparison based on ...	Baringhaus and Franz (2010)
Liu and Modarres (2011)	✗	✓	✗	✓	✓	✗	✓	✗	✓	NA	Independent of p	✗	NA	NA	✓	✓	✓	NA	(✓)	NA	8	1	5	5	Comparison based on ...	Comparison based on ...	Liu and Modarres (2011)
Biswas and Ghosh (2014)	✗	✓	✗	✓	✓	✓	✓	✗	✓	NA	NA	✗	✓	NA	✓	✓	✓	NA	✓	NA	11	0	4	4	Comparison based on ...	Comparison based on ...	Biswas and Ghosh (2014)

Summary and Outlook

Summary and Outlook

- Compared 114 methods based on 20 criteria
- Developed online tool which allows custom filtering and sorting
- Currently working on empirical comparison of top performing methods from theoretical comparison
- Incorporation of data similarity methods into current work on comparison of parametric and Plasmode simulation planned



Thank you for your attention!

References I

[1] B. Aslan and G. Zech. New test for the multivariate two-sample problem based on the concept of minimum energy. *Journal of Statistical Computation and Simulation*, 75(2):109–119, Feb. 2005.

[2] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, Jan. 2004.

[3] M. Biswas, M. Mukhopadhyay, and A. K. Ghosh. A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika*, 101(4):913–926, Dec. 2014.

[4] N. Deb, B. B. Bhattacharya, and B. Sen. Efficiency Lower Bounds for Distribution-Free Hotelling-Type Two-Sample Tests Based on Optimal Transport, Aug. 2021. arXiv:2104.01986 [math, stat].

[5] J. Friedman. On multivariate goodness-of-fit and two-sample testing. Technical report, SLAC National Accelerator Lab., Menlo Park, CA (United States), 2004.

[6] J. H. Friedman and L. C. Rafsky. Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. *The Annals of Statistics*, 7(4):697–717, 1979.

[7] J. H. Friedman and S. Steppel. A nonparametric procedure for comparing multivariate point sets. *Stanford Linear Accelerator Center Computation Research Group Technical Memo*, (153), 1973.

[8] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

[9] P. Ghosal and B. Sen. Multivariate Ranks and Quantiles using Optimal Transport: Consistency, Rates, and Nonparametric Testing, May 2021. arXiv:1905.05340 [math, stat].

References II

[10] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A Kernel Method for the Two-Sample-Problem. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[11] P. Hall and N. Tajvidi. Permutation Tests for Equality of Distributions in High-Dimensional Settings. *Biometrika*, 89(2):359–374, 2002.

[12] N. Henze and B. Voigt. Almost Sure Convergence of Certain Slowly Changing Symmetric One- and Multi-Sample Statistics. *The Annals of Probability*, 20(2):1086–1098, Apr. 1992. Publisher: Institute of Mathematical Statistics.

[13] Z. Huang and B. Sen. A Kernel Measure of Dissimilarity between M Distributions. *Journal of the American Statistical Association*, 0(0):1–27, 2023.

[14] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, Mar. 1951.

[15] F. Liese and I. Vajda. On Divergences and Informations in Statistics and Information Theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, Oct. 2006.

[16] D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

[17] J.-F. Maa, D. K. Pearl, and R. Bartoszyński. Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *The Annals of Statistics*, 24(3):1069–1074, June 1996.

[18] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, June 2017.

References III

[19] S. Mukherjee, D. Agarwal, N. R. Zhang, and B. B. Bhattacharya. Distribution-Free Multisample Tests Based on Optimal Matchings With Applications to Single Cell Genomics. *Journal of the American Statistical Association*, 117(538):627–638, Apr. 2022.

[20] A. Müller. Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability*, 29(2):429–443, June 1997.

[21] P. R. Rosenbaum. An Exact Distribution-Free Test Comparing Two Multivariate Distributions Based on Adjacency. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(4):515–530, 2005.

[22] M. F. Schilling. Multivariate Two-Sample Tests Based on Nearest Neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986.

[23] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *21st Annual Conference on Learning Theory (COLT 2008)*, pages 111–122. Omnipress, 2008.

[24] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010.

[25] G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

[26] G. J. Székely and M. L. Rizzo. The Energy of Data. *Annual Review of Statistics and Its Application*, 4(1):447–479, 2017.

[27] G. Zech and B. Aslan. A new test for the multivariate two-sample problem based on the concept of minimum energy, Sept. 2003. arXiv:math/0309164 version: 1.

Graph definitions

- Nearest neighbor graph: directed graph where each point is connected to its nearest neighbor
- Minimum spanning tree: acyclic graph connecting all points such that the sum of the edge weights (= distances between points) is minimal
- Optimal non-bipartite matching: graph where each point is connected to a single other point such that the sum of the edge weights (= distances between points) is minimal
  Assumption: number of points is even, otherwise delete one point in a way that the matching of the remaining points is optimal

Comparison of cumulative distribution functions, density functions or characteristic functions and methods based on multivariate ranks

Comparison of CDF, density functions or characteristic functions:

- Each distribution is fully characterized by these functions
- Obvious to compare distributions by one of these functions
- Empirical versions of the functions used

Methods based on multivariate ranks

- Ranks-based methods very popular for comparing univariate distributions
- $\mathbb{R}^p$ does not have a natural ordering for $p > 1 \Rightarrow$ generalization not straightforward, but possible e.g. via optimal transport [9, 4]

Kernel-based methods

- Extend feature maps as used by other kernel methods like support vector machines to the space of probability distributions by representing each distribution as a so-called mean function
- This maps each probability distribution to an element in the reproducing kernel Hilbert space (RKHS) corresponding on the chosen kernel
- For characteristic kernels, the distance of the elements in the RKHS is zero iff the distributions coincide [8, 23, 24]
- Example: Maximum mean discrepancy (MMD) [18, 10]: distance of the mean functions measured in the RKHS

Methods based on binary classification

- Idea: use binary classification method trained on the dataset affiliation of each point in the pooled sample
- If the datasets are different, the classifier should be able to distinguish between them, otherwise its performance should be close to random guessing
- Examples:
  - Compare univariate distributions of scores produced by the classifier, e.g. predicted probabilities [5]
  - Classifier two-sample test: uses accuracy of classifier [16]

Others

Distance and similarity measures for datasets:

- Might include properties that are only indirectly captured by the distribution
- Mainly used in meta-learning

Comparison based on summary statistics:

- Comparison of summaries might be less complex than comparison of the datasets themselves

Different testing approaches:

- Test statistic of each two-sample test can be used
- Class contains statistics that cannot be classified into any of the remaining classes

Kernel Measure of Multi-Sample Dissimilarity (KMD)

- Denote the dataset membership of each point in the pooled sample $\{Z_1, \ldots, Z_N\}$ by $\{\Delta_1, \ldots, \Delta_N\}$
- $\{(\Delta_i, Z_i)\}_{i=1}^N$ can approximately be seen as an i.i.d. sample from $(\tilde{\Delta}, \tilde{Z})$ with distribution μ specified by $\mathrm{P}(\tilde{\Delta} = i) = \pi_i, i = 1, \ldots, M$ and $\tilde{Z} | \tilde{\Delta} = i \sim F_i$
- Let $(\tilde{Z}_1, \tilde{\Delta}_1), (\tilde{Z}_2, \tilde{\Delta}_2)$ i.i.d. samples from μ and $(\tilde{Z}, \tilde{\Delta}), (\tilde{Z}, \tilde{\Delta}') \sim \mu$ with $\tilde{\Delta}, \tilde{\Delta}'$ conditionally independent given $\tilde{Z}$
- Denote by K a kernel function over the space $\{1, \ldots, k\}$, e.g. the discrete kernel $K(x, y) := \mathbb{1}(x = y)$
- Then the *kernel measure of multi-sample dissimilarity* (KMD) is defined as

$$\eta(P_1, \ldots, P_k) := \frac{\mathrm{E}\left[K(\tilde{\Delta}, \tilde{\Delta}')\right] - \mathrm{E}\left[K(\tilde{\Delta}_1, \tilde{\Delta}_2)\right]}{\mathrm{E}\left[K(\tilde{\Delta}, \tilde{\Delta})\right] - \mathrm{E}\left[K(\tilde{\Delta}_1, \tilde{\Delta}_2)\right]}.$$

Mukherjee et al. (2022)

- Generalization of the test by Rosenbaum [21] to the *k*-sample problem
- Construct optimal non-bipartite matching on the pooled sample
- Calculate matrix of cross-match counts: each entry is given by the number of matches with one observation coming from one sample and the other from another sample for each pair of samples
- Statistic: Mahalnobis distance of observed cross-counts

Biswas et al. (2014)

- Based on shortest Hamiltonian path (path that visits each vertex exactly once) based on the Euclidean distance
- Statistic: Number of edges connecting points from different datasets + 1

Friedman and Rafsky (1979)

- Based on minimal spanning tree
- Statistic: Number of edges connecting points from different datasets

Rosenbaum (2005)

- Based on optimal non-bipartite matching
- Known as cross-match test
- Statistic: Number of edges connecting points from different datasets

Energy statistic [27]

- Energy statistic [27, 25] is equivalent to Cramér statistic [2]
- e-distance $e(\mathcal{X}, \mathcal{Y})$ between disjoint nonempty subsets $\mathcal{X} = \{X_1, \ldots, X_{n_1}\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_{n_2}\}$ of $\mathbb{R}^p$ is defined as

  $$e(\mathcal{X}, \mathcal{Y}) = \frac{n_1 n_2}{n_1 + n_2} \left(\frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} ||X_i - Y_j||_2 - \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} ||X_i - X_j||_2 - \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} ||Y_i - Y_j||_2 \right)$$

  with $|| \cdot ||_2$ denoting the Euclidean norm.
- The k-sample energy statistic is given by the sum of the e-distances for all $k(k-1)/2$ pairs of samples
- Can be used in bootstrap test procedure for k-sample problem