scientific reports

OPEN

Kendall transformation brings a robust categorical representation of ordinal data

Miron Bartosz Kursa

Kendall transformation is a conversion of an ordered feature into a vector of pairwise order relations between individual values. This way, it preserves ranking of observations and represents it in a categorical form. Such transformation allows for generalisation of methods requiring strictly categorical input, especially in the limit of small number of observations, when quantisation becomes problematic. In particular, many approaches of information theory can be directly applied to Kendall-transformed continuous data without relying on differential entropy or any additional parameters. Moreover, by filtering information to this contained in ranking, Kendall transformation leads to a better robustness at a reasonable cost of dropping sophisticated interactions which are anyhow unlikely to be correctly estimated. In bivariate analysis, Kendall transformation can be related to popular non-parametric methods, showing the soundness of the approach. The paper also demonstrates its efficiency in multivariate problems, as well as provides an example analysis of a real-world data.

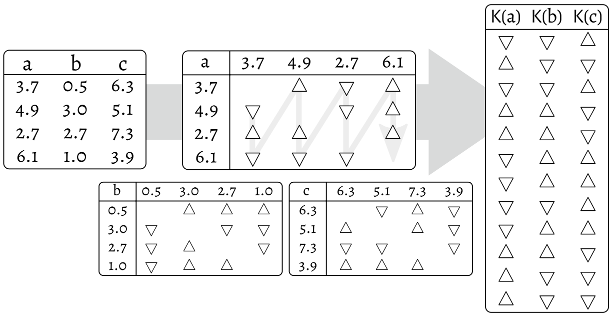
Information theory[1] is a powerful framework utilised in many branches of statistics and machine learning. It is used, among others, for association testing[2], feature selection[3], network reconstruction[4] and clustering[5]. The known deficiency of the theory, though, is that it is well defined for discrete probability distributions[6], yet there is no unequivocally proper generalisation over continuous distributions which would retain important properties without causing substantial theoretical or practical issues.

A common approach here is to simply quantise data before information-theoretic analysis, that is approximate using a small set of discrete values, and treat as categorical afterwards. This is often done through binning, which involves splitting the range of the original feature into consecutive intervals, bins, and then mapping each original value to a discrete one common to a corresponding bin. In a special case of dichotomisation only two bins are used, defined by a single threshold value, which yields a binary outcome. This approach is popular in biomedical applications, as it generates straightforward, comprehensible results[7]. On the other hand, quantisation is a lossy procedure, as it wipes out all intra-bin interactions; consequently it also exaggerates inter-bin differences, which may lead to spurious effects like banding or tautological conclusions. The most substantial problem in context of this work, however, is that quantisation can be done in numerous ways, bringing additional burden of heuristics and hyper-parameters which can critically influence outcomes, especially in the small data limit[8].

The other substantial approach is differential entropy, which is a straightforward generalisation over continuous probability distributions. Unfortunately, while very useful on its own, it violates many useful properties of discrete Shannon entropy and related quantities[9], seriously impacting applicability of more sophisticated methods developed for discrete cases. Moreover, on a practical side, this approach requires modelling and estimation of distributions underlying analysed data, which may easily become a highly non-trivial task relying on cumbersome heuristics, indifferent from quantisation. Henceforth, it is desirable to look for more robust and generic approaches.

Following the idea behind Kendall correlation[10,11], a continuous variable can be represented as a graph of order relations between each pair of its values, and to measure association of variables in terms of their similarity in such representation. Still, the said graph can be thought of not just as an intermediate artefact, but an useful concept on its own. In particular, the list of its edges, when ordered in some arbitrary yet common way, can be written as a discrete variable with at most three states: less, greater and equal. I will refer to conversion to such a form as *Kendall transformation*. This approach involves neither strong assumptions, parameters nor optimisation, and its output can be directly fed into any method expecting categorical input without any adaptations.

Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland.
email: M.Kursa@icm.edu.pl



a	b	c
3.7	0.5	6.3
4.9	3.0	5.1
2.7	2.7	7.3
6.1	1.0	3.9

a	3.7	4.9	2.7	6.1
3.7		△	▽	△
4.9	▽		▽	△
2.7	△	△		△
6.1	▽	▽	▽	

b	0.5	3.0	2.7	1.0
0.5		△	△	△
3.0	▽		▽	▽
2.7	▽	△		▽
1.0	▽	△	△	

c	6.3	5.1	7.3	3.9
6.3		▽	△	▽
5.1	△		△	▽
7.3	▽	▽		▽
3.9	△	△	△	

K(a)	K(b)	K(c)
▽	▽	△
△	▽	▽
▽	▽	△
△	△	▽
△	△	▽
▽	△	△
▽	△	△
▽	▽	△
▽	△	△
△	△	▽
△	▽	▽
△	▽	▽

Figure 1. Kendall transformation of a toy information system with four observations (left). Three continuous features, *a*, *b* and *c*, are converted into matrices of relations between values (centre). Matrices are flattened in an arbitrary but consistent manner and collected into a transformed system, which consists of twelve observations, but only has categorical values (right).

Although it only preserves the ranking of the original variable, the success of rank tests[12] suggests it is an effective trade-off in many practical applications.

In this paper I will analyse properties of Kendall transformation and argue that it is a sufficient and reliable representation for many information-based inquiries into the data, consequently is a robust and effective alternative to both binning and elaborate continuous entropy estimators.

Kendall transformation

Let us assume some way of ordering all $m = n(n-1)$ possible pairs from a set $\{1..n\}$, with $(a_j, b_j \neq a_j)$ denoting j-th such pair. Without a loss of generality, I will assume it to be one corresponding with a column-wise traversal of an $n \times n$ matrix. Furthermore, let us assume X to a be a totally ordered set, that is equipped with a greater-or-equal relation $\geq$ that is reflexive, transitive, antisymmetric and strongly connected.

For an arbitrary n-element vector $x \in X^n$, Kendall transformation is a function $\mathsf{K} : X^n \rightarrow \{\triangle, \triangledown, \square\}^m$ defined as

$$\mathsf{K}(x)_j := \begin{cases} \triangle \text{ if } x_{a_j} < x_{b_j} \\ \triangledown \text{ if } x_{a_j} > x_{b_j} \\ \square \text{ if } x_{a_j} = x_{b_j} \end{cases}. \tag{1}$$

By extension, Kendall transformation of an information system exclusively composed of totally ordered features is an information system in which every feature has been Kendall-transformed. Figure 1 presents an illustrative example of transformation of a small, toy system. It is trivial to note that $\mathsf{K}(x) = \mathsf{K}(f(x))$ for any strictly increasing f, as $f(a) \geq f(b) \rightarrow a \geq b$ in this case. Henceforth, the proposed transformation is lossy, but exactly preserves observation ranking.

Bivariate analysis.

Let us further assume for a moment that there are no ties between values, hence that the equal state does not occur. With two features x and y, we say that pair j is concordant if $\mathsf{K}(x)_j = \mathsf{K}(y)_j$ and discordant otherwise. The Kendall correlation coefficient τ is then the difference between the number of concordant and discordant pairs, normalised by m (One should note that τ is quantised into $m/2 + 1$ states; in particular, $\tau = 0$ is only possible for n or $n-1$ divisible by 4). Furthermore, the entropy of a Kendall-transformed vector is $\log(2)$, and the mutual information (MI) between two transformed variables I^{K} is a function of a Kendall correlation τ between them, namely

$$I^{\mathsf{K}}(\tau) = \tau \log \sqrt{\frac{1+\tau}{1-\tau}} + \log \sqrt{1-\tau^2}. \tag{2}$$

It is an even function, strictly increasing for $\tau > 0$. Moreover, it can be also considered as an extension of a simple, commonly used formula connecting MI and correlation coefficient ρ,

$$I^G(\rho) = -\log \sqrt{1-\rho^2}. \tag{3}$$

It is derived by subtracting differential entropy of a bivariate normal distribution from a sum of differential entropies of its marginals, with a constant factor $\log \sqrt{2\pi e}$ omitted, and, although valid only for the Pearson correlation, it is often used with other coefficients, usually Spearman's[13]. Both functions behave similarly for small absolute values of τ and ρ. On the other hand, when respective correlation coefficient approaches 1 or − 1, I^{K} achieves maximum, while I^G diverges, causing problems in certain use cases, especially when highly correlated features are of interest.

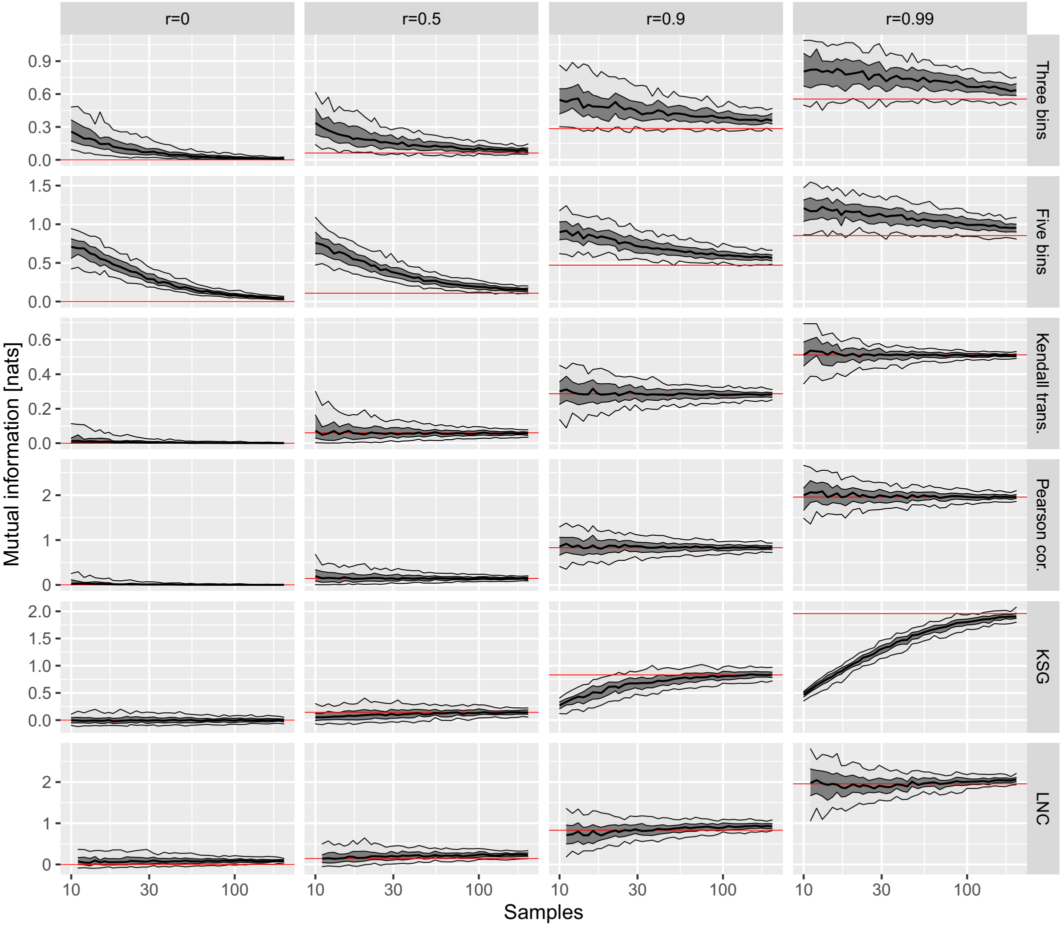


Figure 2. Mutual information between two variables drawn from a bivariate normal distribution with correlation r, calculated with different methods. The bands show 5th, 25th, 50th, 75th and 95th percentile over 100 realisations, while red lines the theoretical value, or its estimate for 5000 samples.

As with Pearson or Spearman[11,14] correlation, on Kendall-transformed information systems we can only detect monotonic relationships; if, say, $x \sim \mathscr{U}(-1, 1)$, the relation between x and x^2 will be lost. I will argue that it is well justified constraint in small-n problems, however. Under the null hypothesis of interaction testing, independence between variables, the probability of each joint state is a product of its marginal probabilities. It is very unlikely to get a small sample of such symmetry, though, hence the agreement with null may easily become less likely than with the alternative hypothesis. Obviously, we may apply corrections for this phenomenon, but for a cost of severely hindered sensitivity. On the other hand, even among short sequences, the probability of randomly getting a sorted one is minimal ($1/n!$ without ties), hence the null of non-monotonicity is much more robust.

Let us empirically investigate the properties of I^{K} on a synthetic data drawn from bivariate normal distribution; we set the marginal distributions to $\mathscr{N}(0, 1)$, and covariance to r. For a comparison, we will use five other methods of estimating MI: quantisation into three or five bins of equal width, Pearson correlation coefficient with Eq (3), finally two versions of the k-nearest neighbour estimator, variant 2 of the Kraskov estimator[15] with $k = 5$ (KSG) and its extension corrected for local non-uniformity[16] with $k = 10$ and $\alpha = 0.65$ (LNC). When not noted otherwise, a maximal likelihood estimator of entropy is used. The results of this experiment, repeated for four values of r, various sample sizes n and over 100 random replicates, are collected on Fig. 2.

We can see that quantisation leads to a very poor convergence, which makes such estimators unsuitable for low n cases; this is especially pronounced for $r = 0$, which in practice would lead to many false positives in association sweeps. In contrast, I^{K} provides nearly unbiased results even for smallest values of n, as well as a relatively small variance.

The remaining methods estimate differential MI, so $-\log\sqrt{1 - r^2}$. The one based on Pearson correlation, which behaviour closely resembles this of I^{K}, yet with a smaller variance. Still, this is due to the fact that we have fully satisfied strong assumptions this estimator relies on; it won't be nearly as effective in a general case. The KSG estimator has a very low variance, and provides a bell-shaped distribution for estimates of $r = 0$, which is handy for independence testing. Still, in highly correlated cases it exhibits its known deficiency of a very slow convergence. The LNC estimator, on the other hand, converges fast for any r; it is visibly biased in the independence case, however, which is likely to hurt specificity.

Clearly, Kendall transformation is the most versatile and robust among investigated solutions; it works reliably over the entire range of analysed cases, takes no parameters and is never substantially inferior to the best method.

Ties. As mentioned earlier, the entropy of a Kendall-transformed variable is log(2) if there are no ties, regardless of the distribution of the original. This becomes intuitive given that this transformation, similarly to ranking, wipes scale information and retains only order; hence, it effectively converts any input distribution into an uniform one. Tied values cannot be separated by a such process, hence the resulting effective distribution in a general case is a mixture of an uniform distribution and Dirac deltas located over tied values, which is a richer, more complex structure. In a similar fashion, the introduction of ties generate □ states in the transformed variable which first increases its entropy, up to log(3), when the complexity contributions of uniform and discrete components become balanced. Additional ties effectively convert the distribution into a discrete one, decreasing the entropy of the transformation. Finally, due to coalescence of values, only one state remains both in the original variable and its Kendall transformation, and entropies of both become 0.

Naturally, the correspondence between actual entropy of a discrete variable and the entropy of a Kendall transformation of its numerical encoding holds only in a constant and binary case; otherwise the order in which states are encoded becomes important. Henceforth, Kendall transformation is directly applicable to numeric, ordinal and binary features, and can provide a viable representation of ties when they are not numerous enough to dominate the ordinal nature of a feature. The proper handling of arbitrary categorical data in such a framework is a subject for further research, though, by a simple extension, we may analyse such features by breaking them into a set of category-vs-other indicator features.

One should note, though, that the above reasoning applies to actual ties, i.e., pairs of values that are indistinguishable, like two days without rain; as opposed to two days with such a similar amount of precipitation that the resolution of the sensor is insufficient to differentiate them. In the latter case it is better to break such artificial ties using random jitter or to treat comparisons between them within the Kendall-transformed variable as missing observations.

Interestingly, when x is ordered and contains no ties and y is binary, I^{K} also corresponds to a popular measure of association. Namely, it is a function of A defined as the area under the receiver operating characteristics curve[17] (AUROC),

$$I^{\mathsf{K}}(A;a;b)=\frac{2ab}{n(n-1)}\left(A\log\frac{A}{1-A}+\log(2-2A)\right), \tag{4}$$

normalised by sizes of both classes, a and $b=n-a$. This way, Kendall transformation is also connected with the Mann-Whitney-Wilcoxon test[18], as its statistic $U=ab(1-A)$.

Multivariate analysis. The important gain of Kendall transformation is that transformed features can be used in more complex considerations that just bivariate ones. In particular, we can calculate joint, conditional or multivariate mutual information and use it to investigate relationships between features.

For example of such an analysis, let us consider an information system composed of three independent, random features $a,b,c\sim U(0,1)$ and a decision which is either a linear $y=a\lambda+b(1-\lambda)$ or an example non-linear $y=\max(a\lambda,b(1-\lambda))$ function of a and b. It is worth noticing that although linear relation seems pretty basic, Kendall transformation makes it invariant to monotonic transformations of a, b and y, hence this case covers more complex functions, for instance $y=\sinh^{\mu}(a)\cdot(b-b_0)^{\nu}$. Let us also denote the Kendall-transformed features with upper-case letters, i.e., $A=\mathsf{K}(a)$ and so on. Figure 3 contains the maximum likelihood estimates of certain information scores in such system, for $n=200$ and a selection of realisations for a range of λ values.

We can see that for most realisations the joint mutual information $I(A,B;Y)$ is larger than either marginal mutual information, signifying that considering both features allows for a better prediction of Y. Their interaction can be directly measured with conditional mutual information $I(A;B|Y)$; we can also compare it to a baseline of $I(A;C|Y)$, which is asymptotically zero as C is irrelevant. This score substantially dominates baseline for almost full range of λ in the linear case, and at least half of it in the much harder non-linear case, confirming the previous conclusion of the presence of interaction. Moreover, it reaches maximum for $\lambda=1/2$, i.e., balanced impacts of a and b, and decreases as either of them dominates, which is a reasonable outcome.

In a yet another view, we can analyse the three-feature mutual information $I(A;B;Y)$. Because of the independence of A and B, it is approximately equal to $-I(A;B|Y)$, hence has a negative minimum for $\lambda=0$, which also signifies synergy between the three features. Similarly, the marginal MIs $I(A;Y)$ and $I(B;Y)$ behave as a reasonable measure of relative impact, akin weights in linear regression, yet generalisable over many non-linear relationships.

The other important aspect of using Kendall transformation for multivariate analysis is how well we can estimate the joint distribution, which generally depends on how many states are there, and how many observations of each can we count. The number of states increases exponentially with the dimension of the analysis, yet the base of said exponent is usually 2 when using Kendall transformation and $5-10$ in case of typical quantisation. Also, due to the nature of the Kendall transformation, the effective number of observations is squared and marginal distributions are well balanced, reducing the noise in counts and the risk of spuriously unobserved states.

Kendall-transformed information systems can also be used as an input for sophisticated machine learning methods. Some approaches can operate in a straightforward way, like many feature selection methods which may be used for explanatory analysis, others may require appropriate adaptations. The particular caveat is that the transformation generates artificial correlation structure between pairs, hence methods which rely on them being independent may become biased.

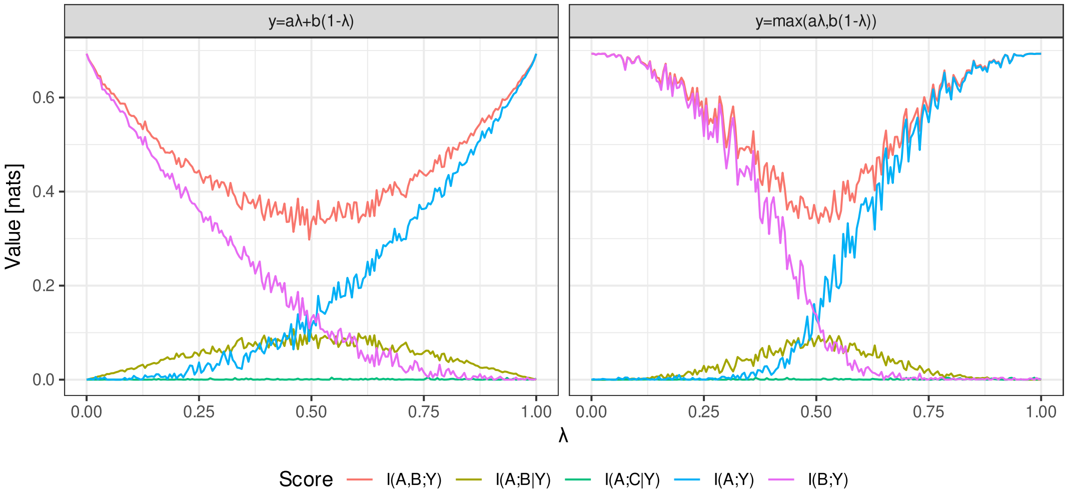


Figure 3. Values of certain information scores for a Kendall-transformed system of three features a, b and y engaged in a simple interaction, $y = a\lambda + b(1-\lambda)$ (left) or $y = \max\{a\lambda, b(1-\lambda)\}$ (right), for a range of realisations with a different λ parameter.

Inverse transformation. A Kendall-transformed feature can be transformed back using the following method. First, for each object i we extract values of pairs $(i, \cdot)$, and score each $\triangle$ with 1 and each $\triangledown$ with -1. Then, we order objects according to the decreasing score, which recovers the ranking. Such algorithm is equivalent to the Copeland's election method[19], and it is trivial to see that it always correctly reverses a valid result of a Kendall transformation.

Kendall transformation is non-surjective, henceforth, when operating on transformed data, we can stumble upon invalid sequences of $\triangle$, $\triangledown$ and $\square$. For instance, let's consider a feature $(\triangle, \triangledown, \triangledown, \triangle, \triangle, \triangledown)$ — it corresponds to a cycle of three observations, and cannot be generated by transforming any ranking. While there is no universal solution covering such cases, Copeland's method can be easily adapted and used as one of possible heuristic solutions; in this particular case, it maps these features into all observations with the same rank, which intuitively seems as a reasonable solution.

Furthermore, sometimes we can also expect missing values or fuzzy predictions, i.e., probabilities or weights of each state. In this cases, we can reach out deeper into the social choice theory for a more sophisticated approach, like the Schulze's[20] or Tideman's[21] method. They will produce a consensus graph of ordering relations, from which we can recover complex tie structures and disconnected subsets of observations.

Merging of transformed data. One of the common problems in molecular biology is the lack of consistent calibration of certain experimental procedures, especially the high-throughput ones. From a mathematical point of view, this phenomenon can be well modelled as an assumption that results for each batch of samples are filtered by an arbitrary monotonic function, specific to a given batch. This way, explanatory outcomes from different batches are mostly consistent; yet, raw datasets cannot be naively combined, for instance to look for more subtle aspects in meta-analysis. It is because the batch effect can be very substantial, even to the point of overshadowing actual interactions, heavily biasing the results.

Kendall transformation can be used as another approach to tackle this problem. The independently transformed data sets from different batches do not retain the strictly increasing dis-calibration effects, hence they can be safely combined without the risk of introducing bias. The merged data will not represent relations in cross-dataset pairs, but we can treat these rows as missing and still conduct actions like feature selection or model training. Only for the prediction, the sets have to be separated to perform inverse transformation—the cross-set relations will be left unknown, but so they were in the first place.

Example

For an example analysis using the Kendall transformation, let us consider the morphine withdrawal data set from[22]. It collects concentrations of 15 neurotransmitters in 6 brain structures, measured in four groups of rats: subject to either a morphine or saline treatment, as well as measured directly after treatment or after a 14 day withdrawal period and re-exposure to the administration context. Additionally, ultrasonic vocalisation (USV) intensity of rats was also quantified as a number of episodes during a 20 minute recording session. In contrast to the original, in the data used here one record has been rejected due to missing data.

Overall, the set is composed of 37 objects and 90 continuous features, distributions of which are predominantly not normal, as well as three separate decision features, continuous USV episode count, and two categorical: morphine treatment and withdrawal period. Such structure is typical to many biomedical studies.

Feature ranking. In the original paper, a standard, bivariate non-parametric statistical analysis was used to identify compounds significantly connected with each of the decision features. Namely, categorical decisions

	USV	Morphine	Withdrawal
Kendall transformation	1.00	1.00	1.00
Three equal-width bins	0.71	0.50	0.56
Five equal-width bins	0.83	0.50	0.62
Three equal-frequency bins	0.44	1.00	0.86
Five equal-frequency bins	0.40	0.17	0.64
Random Forest importance	0.80	1.00	0.86

Table 1. Agreement of mutual information feature rankings obtained using different data transformations with significant results of a standard non-parametric statistical analysis applied to the morphine withdrawal data set, given as a maximal value of the Jaccard index. Random Forest importance ranking added for comparison.

	USV [Spearman cc.]	Morphine [AUROC]	Withdrawal [AUROC]
Kendall-transformed	60.1% [46.1–70.1%]	82.8% [71.4–89.3%]	100.0% [98.1–100.0%]
Original data	62.2% [49.2–69.7%]	87.1% [76.5–96.8%]	100.0% [97.4–100.0%]

Table 2. Test-set accuracy of a Random Forest classifier trained on the morphine withdrawal data set, Kendall-transformed and original. Values provided as median [IQR] over 100 simulation realisations.

were analysed with Mann-Whitney-Wilcoxon U test[18], while continuous one with a Spearman ρ test[23]. Such analysis, combined with a multiple comparisons correction, yields 5 significant features for the episode count problem, 1 for morphine and 13 for withdrawal. Let us compare such outcome with the mutual information rankings of features obtained on either Kendall-transformed or binned data, as well as with the Random Forest importance[24] applied directly (using the randomForest R package[25], in its default set-up). Their agreement can be quantified by the maximal value of Jaccard index over all possible cut-offs in the respective ranking. Precisely, Jaccard index[26,27] measures similarity of two sets, A and B, and is defined as

$$J(A, B) := \frac{|A \cap B|}{|A \cup B|}. \tag{5}$$

Here, we have a set of all features, Ω, a golden standard set $G \subset \Omega$ of features selected by a standard analysis, and a ranking or scoring of features which can be commonly expressed as a function $s : \Omega \to \mathbb{R}$, having greater values for more important features. Then, the final score can be formally defined as

$$J_{max}(G, s) := \max_{t \in \mathbb{R}} J(G, \{x \in \Omega : s(x) > t\}). \tag{6}$$

Table 1 collects the results of the said experiment; we see a perfect agreement in case of Kendall transformation, which is unsurprising given the aforementioned equivalence relations, as well as the fact that Spearman and Kendall correlation coefficients are usually highly correlated. The rankings on quantised data are substantially influenced by the binning method; even though all predictors are the same, a different method is optimal for each decision. Perfect agreement with baseline is only achieved once, and the agreement is pretty poor on average. The much more elaborated approach, RF importance, achieves a relatively high average agreement of 0.89, given that it also considers multivariate relations. Overall, the results support the notion that simple binning is susceptible to exaggerating spurious interactions inherent to small sample data. Kendall transformation not only helps to avoid this phenomenon, but also requires no hyper-parameters, which have a critical impact on binning.

Prediction. In order to investigate Kendall transformation in the context of machine learning, let us now compare the accuracy of the Random Forest method applied to the morphine data directly and after Kendall transformation. Accuracy is estimated using a bootstrap train-test split, hence with roughly 63% of observations used to build a model predicting the remaining data—in both cases the split is done before transformation. For the modelling after Kendall transformation, predictions for pairs are translated to ranking of original objects using the Copeland's method proposed in Section Inverse transformation, considering the fraction of votes cast for the △ state. Similarly, for the baseline, the vote fraction is used as an outcome for classification problems (Morphine and Withdrawal), while the regression mode is used for the USV count prediction. Finally, the accuracy is quantified with Spearman correlation, for regression, or with AUROC, for classification. As in previous example, the randomForest[25] implementation is used.

The result of this analysis, repeated over 100 random replications, is reported on Table 2. We see that the application of Kendall transformation, despite its lossy nature, does not substantially hinder the accuracy of the prediction; the loss is statistically significant only in case of the Morphine problem, yet it is about a quarter of the interquartile range. Thus, we can conclude that Kendall-transformed data is a viable input for a machine learning method.

	USV	Morphine	Withdrawal
Kendall transformation	95.1% [93.5–96.0%]	96.0% [93.8–96.9%]	98.3% [97.3–98.7%]
Naive merge	26.6% [24.5–39.4%]	41.8% [-23.8–47.5%]	71.7% [21.6–91.2%]

Table 3. Robustness of Kendall transformation-based feature ranking to a simulated loss of calibration between two measurement sessions, quantified by Spearman correlation with the ranking on unperturbed data, using the morphine withdrawal example. Values provided as median [IQR] over 100 simulation realisations.

Data integration. For the sake of this example, let us simulate the acquisition of data from two incoherent measurements. To this end, the morphine set is randomly split in half, and the values of features in one part is tripled. Next, we either naively merge halves back and apply Kendall transformation on the fused set, or the other way round, merge independently transformed parts. Finally, we compare the agreement of mutual information-based feature rankings obtained after each processing with this calculated on the golden standard, Kendall-transformed original, unperturbed data.

The results, quantified by Spearman correlation coefficient and averaged over 100 realisations, are presented as Table 3. We see that the simulated measurement artefact has a substantial, negative impact on the accuracy and stability of the ranking in the naive scenario, while the merge of Kendall-transformed parts reliably achieves over 90% agreement. This is expected due to a fact that the applied disruption is a strictly increasing function, which Kendall transformation is invariant to. The agreement is not perfect, however, because the set merged after transformation lacks almost 50% of information, namely all the pairs representing intra-half relationships—in light of this fact, the observed loss is rather limited.

Conclusions

Kendall transformation is a novel way to represent ordinal data in a categorical form, as well as to apply discrete methods and approaches on such data. While standard quantisation procedures sacrifice precision, Kendall transformation precisely preserves ranking, sacrificing original distribution instead. This approach is common to non-parametric statistical methods, though, and it is proved effective in many use-cases, in particular in small sample size conditions. In fact, I show that Kendall transformation is tightly connected with certain popular methods of this class on a theoretical level.

Moreover, Kendall transformation is reversible into ranking, has no parameters and imposes no restrictions on the input, as well as offers consistent behaviour regardless of its characteristics. The method is also very versatile, as its output can be used both to calculate some simple coefficient and be a part of an elaborate algorithm.

Data availability

The implementation of Kendall transform (and its inverse) for R is available in CRAN package praznik[28], starting from version 8.0. The code used to execute the computational experiments presented in the paper can be found on GitHub, https://github.com/mbq-suppl/kendall-transformation.

Received: 3 January 2022; Accepted: 29 April 2022
Published online: 18 May 2022

References

1. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Techn. J.* **27**, 379–423 (1948).
2. Smith, R. A mutual information approach to calculating nonlinearity. *Stat* **4**, 291–303 (2015).
3. Brown, G., Pocock, A., Zhao, M.-J. & Lujan, M. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **13**, 27–66 (2012).
4. Margolin, A. A. *et al.* ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**, 1–15 (2006).
5. Brown, P. F., De Souza, P. V., Mercer, R. L., Pietra, V. J. D. & Lai, J. C. Class-based n-gram models of natural language. *Comput. Linguist.* **18**, 467–479 (1992).
6. Seok, J. & Kang, Y. S. Mutual information between discrete variables with many categories using recursive adaptive partitioning. *Sci. Rep.*5 (2015).
7. Altman, D. G. & Royston, P. The cost of dichotomising continuous variables. *BMJ* **332**, 1080 (2006).
8. Budden, D. & Crampin, E. J. Information theoretic approaches for inference of biological networks from continuous-valued data. *BMC Systems Biology***10** (2016).
9. Jaynes, E. Information theory and statistical mechanics. In Ford, K. (ed.) *Brandeis University Summer Institute Lectures in Theoretical Physics*, vol. 3, pp 181–218 (W.A. Benjamin).
10. Kendall, M. G. A new measure of rank correlation. *Biometrika* **30**, 81–93 (1938).
11. Croux, C. & Dehon, C. Influence functions of the Spearman and Kendall correlation measures. *Stat. Methods Appl.* **19**, 497–515 (2010).
12. Hájek, J., Šidák, Z. & Sen, P. K. *Theory of Rank Tests* 2nd edn. (Academic Press, San Diego, 1999).
13. De Jay, N. *et al.* MRMRe: An R package for parallelized mRMR ensemble feature selection. *Bioinformatics* **29**, 2365–2368 (2013).
14. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **15**, 72 (1904).
15. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat Interdiscip. Topics* **69**, 16 (2004).
16. Gao, S., Ver Steeg, G. & Galstyan, A. Efficient estimation of mutual information for strongly dependent variables. *J. Mach. Learn. Res.* **38**, 277–286 (2015).
17. Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**, 861–874 (2006).

18. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).
19. Copeland, A. H. A 'reasonable' social welfare function (1951).
20. Schulze, M. A new monotonic, clone-independent, reversal symmetric, and Condorcet-consistent single-winner election method. *Soc. Choice Welfare* **36**, 267–303 (2011).
21. Tideman, T. N. Independence of clones as a criterion for voting rules. *Soc. Choice Welfare* **4**, 185–206 (1987).
22. Hamed, A. & Kursa, M. B. Inter-individual differences in serotonin and glutamate co-transmission reflect differentiation in context-induced conditioned 50-kHz USVs response after morphine withdrawal. *Brain Struct. Funct.* **223**, 3149–3167 (2018).
23. van de Wiel, M. & Di Bucchianico, A. Fast computation of the exact null distribution of Spearman's ρ and Page's L statistic for samples with and without ties. *J. Stat. Plan. Inference* **92**, 133–145 (2001).
24. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
25. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News***2**, 18–22 (2002). https://CRAN.R-project.org/doc/Rnews/.
26. Jaccard, P. Article in Bulletin de la Societe Vaudoise des Sciences Naturelles. *Bull. Soc. Vaudoise des Sci. Nat.* **37**, 547–579 (1901).
27. da Fontoura Costa, L. Further generalizations of the Jaccard index. arXiv:2110.09619 (2021).
28. Kursa, M. B. Praznik: High performance information-based feature selection. *SoftwareX* **16**, 100819. https://doi.org/10.1016/j.softx.2021.100819 (2021).

Author contributions

M.B.K. proposed the approach, performed all the computational experiments and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.B.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.