ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Fushing H, Roy T. 2018 Complexity of possibly gapped histogram and analysis of histogram. *R. Soc. open sci.* **5**: 171026.
http://dx.doi.org/10.1098/rsos.171026

Received: 2 August 2017
Accepted: 29 January 2018

Subject Category:
Computer science

Subject Areas:
complexity/statistical physics/computational physics

Keywords:
data mechanics, hierarchical clustering algorithm, macro-state, statistical mechanics, unsupervised machine learning

Author for correspondence:
Hsieh Fushing
e-mail: fhsieh@ucdavis.edu

Complexity of possibly gapped histogram and analysis of histogram

Hsieh Fushing and Tania Roy

Department of Statistics, University of California, Davis, CA 95616, USA

HF, 0000-0002-9292-6980

We demonstrate that gaps and distributional patterns embedded within real-valued measurements are inseparable biological and mechanistic information contents of the system. Such patterns are discovered through data-driven possibly gapped histogram, which further leads to the geometry-based analysis of histogram (ANOHT). Constructing a possibly gapped histogram is a complex problem of statistical mechanics due to the ensemble of candidate histograms being captured by a two-layer Ising model. This construction is also a distinctive problem of Information Theory from the perspective of data compression via uniformity. By defining a Hamiltonian (or energy) as a sum of total coding lengths of boundaries and total decoding errors within bins, this issue of computing the minimum energy macroscopic states is surprisingly resolved by applying the hierarchical clustering algorithm. Thus, a possibly gapped histogram corresponds to a macro-state. And then the first phase of ANOHT is developed for simultaneous comparison of multiple treatments, while the second phase of ANOHT is developed based on classical empirical process theory for a tree-geometry that can check the authenticity of branches of the treatment tree. The well-known *Iris* data are used to illustrate our technical developments. Also, a large baseball pitching dataset and a heavily right-censored divorce data are analysed to showcase the existential gaps and utilities of ANOHT.

1. Introduction

Without spatial and temporal coordinates, a sample of one-dimensional real-valued measurements is generally taken as one basic simple data type and receives very limited research attention. Its simplicity is seemingly implied by the illusion that its information contents are evident and transparent, and all its characteristic patterns should have immediately popped up right in front of our eyes. As a matter of fact, its pictorial representation, usually called an empirical distribution, indeed embraces hidden and implicit patterns waiting to be extracted.

THE ROYAL SOCIETY PUBLISHING

There are many possible patterns that can be exhibited through the piecewise step-function structure of an empirical distribution. Among all possible patterns, two of them take the most basic forms: one is 'linear segment' and the other is 'gap'. A linear segment indicates a potential uniform distribution being embedded within the empirical distribution, and a gap strictly indicates an interval zone, in which definitely allows no observations. As for the rest of the potential patterns, they can be very well approximated by properly combining these two basic patterns. Therefore, an empirical distribution ideally can be well approximated by various serial compositions of basic patterns. Each composition is a possibly gapped piecewise linear approximation, which is correspondingly equivalent to a possibly gapped histogram.

From such an approximation perspective, the larger number of patterns involving in such a series of basic patterns mean a higher cost in terms of data compression [1]. Nevertheless, for the sake of true and intrinsic data patterns, it seems very natural to think that the hypothetical population distribution underlying the observed empirical distribution should embrace discontinuity.

But this is not the case in the statistical literature. Even though a histogram is discrete in nature, most existing versions of its constructions were developed by approximating a density function by imposing continuity and smoothness assumptions [2–4]. In sharp contrast, in computer science literature, a histogram is always discrete because it is primarily used for visualizing data or queries from a database [5]. So it does not involve approximations via continuous components of uniform or other smooth kernel distributions.

The computing costs for a histogram via aforementioned statistics and computer science viewpoints are not severely high. But the computing load for a possibly gapped histogram can be theoretically very heavy because a constructive approach ideally needs to allow an unknown number of bins with heterogeneous bin-width and to accommodate an unknown number of gaps of different sizes. Furthermore, the computing cost will grow with the sample sizes because of the multi-scale nature of these two basic patterns. That is why the construction of a possibly gapped histogram is computationally complex.

In this paper, we resolve this computational issue algorithmically by treating it as if it is derived from a physical system. From this perspective, a possibly gapped histogram should embrace deterministic structures, which are all boundaries of the bins, and the stochastic structures that are uniform within each bin. That is, the deterministic and stochastic structures together constitute the system information contents of a one-dimensional dataset.

Such physical information contents render another interpretation from the perspective of algorithmic complexity [6]. A relatively simple way of seeing this complexity is the fact that the candidate ensemble is constructed via a two-layer Ising model [7]. This ensemble grows exponentially in size with respect to the number of data points, say n. Based on this ensemble description, we clearly see that all boundary parameters are characteristically local because of multiple relevant scales also depending on n in a heterogeneous fashion across all potential bins.

Further, it is interesting to note that the decoding error under uniformity is nearly independent of sample size within each bin. This fact becomes an effective criterion for confirming uniformity, on one hand. On the other hand, any further division of a uniform distribution would give rise to several uniform distributions with lower total decoding errors. Therefore, we need to balance between the total decoding errors and coding lengths of boundaries of the bins (with respect to a converting index). So a Hamiltonian is defined.

Therefore, constructing a possibly gapped histogram is indeed a complex physical problem of statistical mechanics aiming for extracting the lowest Hamiltonian macroscopic state (or macro-state). It is also a problem of information theory aiming for the best balance between the costs. But it is clearly not a problem of statistics because of its multi-scale nature. Thus, we need a brand-new computing protocol to seek for the optimal solution and the macro-state.

It is surprising that this seemingly complex macro-state can be approximated by a relatively simple computational algorithm. By applying the popular hierarchical clustering (HC) algorithm with complete or other modules, excluding the single-linkage one, on a one-dimensional dataset, the resultant hierarchical tree indeed gives rise to just a few feasible candidates. We develop an algorithm to select one from this small set of candidates. Nearly optimal solutions can be derived. Also, the criterion of decoding error of uniformity turns out to be a practical way of checking whether a space between two consecutive bins is indeed a gap.

The merits of a possibly gapped histogram are intriguingly profound and far-reaching. We demonstrate its potential merits through our developments of new data analysis paradigm, called analysis of histogram (ANOHT). The first phase of ANOHT is designed to address the issues: where

and how are multiple treatments or distributions locally and globally different? The second phase of ANOHT is designed to answer issues: which treatments are closer to which, but far away from others? And why? Here ANOHT is simple and equipped with excellent visualization capability. ANOHT also gives the biological and mechanistic meanings to identify gaps found within a histogram.

Our technical developments are illustrated throughout by employing the well-known *Iris* data. In the Results sections, ANOHT is first applied onto a large pitching dataset of three Major League Baseball (MLB) pitchers, and then onto a heavily censored divorce dataset. The implications of our developments from building a possibly gapped histogram to ANOHT are discussed in the Conclusion section.

2. Methods

In this section, our technical developments are divided into four subsections: (i) building the ensemble of candidate histograms; (ii) deriving the decoding errors; (iii) constructing the algorithm for possibly gapped histogram as the first phase of ANOHT; and (iv) developing the tree geometry as the second phase of ANOHT. The *Iris* data are illustrated throughout our developments.

2.1. Candidate ensemble

Let $\{x_i^o\}_{i=1}^n$ be the observed data points, and their ranked values as $\{x_{(i)}^o\}_{i=1}^n$ in increasing order. The construction of a possibly gapped histogram can be precisely stated in the following two-layer one-dimensional Ising model depicted in figure 1.

1. Layer-1: a 'spin' is placed on each spacing between a pair of consecutive observed values $(x_{(i-1)}^o, x_{(i)}^o)$. First an up-spin for sharing the same uniform-part, and then a down-spin for indicating that $(x_{(i-1)}^o, x_{(i)}^o)$ belongs to two distinct bins with or without a gap;
2. Layer-2: a 'hidden-spin' is placed between two consecutive uniform-parts, on the first layer: (1) '+'-spin for having a gap and (2) '−'-spin for without a gap.

In total, there are $n-1$ up- or down-spins in the Layer-1. Then the number of spins in the Layer-2 is exactly the number down-spins, say k, in the Layer-1. Since C_k^{n-1} is the number of distinct sets of k nodes chosen from the pool of $n-1$ nodes, this class of 'two-layer one-dimensional Ising model' has its cardinality growing exponentially with the sample size as:

$$3^{n-1} = \sum_{k=0}^{n-1} C_k^{n-1} 2^k.$$

This exponential growth rate reflects the essence of the fact that all boundaries are local once we revoke continuity and smoothness assumptions. That is, the complexity of computing within this ensemble comes from dealing with multi-scales of parameters, which is one significant data-driven nature of possibly gapped histogram.

2.2. Decoding errors

Next we consider the decoding error coming from uniform data-compression. First let the identically independently distributed (i.i.d.) random variables be denoted as $\{X_i\}_{i=1}^n$, that is, X_i is distributed with respect to (w.r.t.) standard uniform density function $f(x) = F'(x) = 1,\ \forall x \in [0,1]$. Denote the order statistics as $\{X_{(i)}\}_{i=1}^n$, and the density of kth-order statistic $X_{(k)}$, say $g_k(x)$, is evaluated as below.

$$\begin{aligned}\Pr[x < X_{(k)} < x+\delta] &= C_k^n [F(x)]^{k-1} [F(x+\delta) - F(x)][1 - F(x+\delta)]^{n-k} \\ &\approx C_k^n [F(x)]^{k-1} f(x) [1 - F(x+\delta)]^{n-k} \delta \\ &= g_k(x)\delta\end{aligned}$$

so that,

$$g_k(x) = C_k^n x^{k-1} [1-x]^{n-k}.$$

Here $g_k(x)$ becomes the exponential density for the extreme ordered statistics $X_{(1)}$ $(k=1)$ and $X_{(n)}$ $(k=n)$.

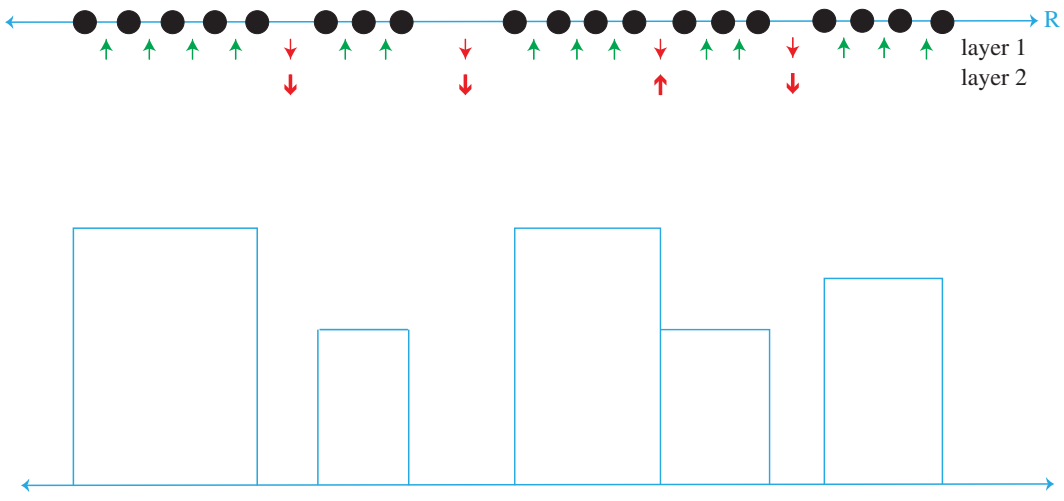

Figure 1. Schematic of two-layer Ising model. The black dots on the blue line represent observations ordered and placed on the real line. The two layers of spins are given below in the next two rows, where up and down arrows describe '+' and '−' spins, respectively. A green arrow means that the two consecutive pairs of observations will be in the same bin and a red arrow means they will be in different bins. Hence, in the second layer, the down red arrow indicates two separate bins with an existential gap, and an upward red arrow indicates two consecutive bins, but with no gap. Next, the histograms are constructed accordingly.

Then we have,

$$E[X_{(k)}] = \frac{k}{n+1}; \quad E[X_{(k)}^2] = \frac{k(k+1)}{(n+1)(n+2)}$$

and

$$\mathrm{Var}[X_{(k)}] = E[X_{(k)}^2] - \{E[X_{(k)}]\}^2 = \frac{k(n-k+1)}{(n+1)^2(n+2)}.$$

Also, we have,

$$\sum_{k=1}^{n} \mathrm{Var}[X_{(k)}] = \frac{n}{6(n+1)}.$$

Given the observed data as $\{x_{(i)}^o\}_{i=1}^n$, further let $\tilde{X}_i \sim U[0,1]$ also be another set of observations, distributed with standard uniform distribution. Then decoding error sum of squares (DESS) is evaluated as

$$\begin{aligned}\sum_{k=1}^{n}[\tilde{X}_{(k)} - x_{(k)}^o]^2 &= \sum_{k=1}^{n}[\tilde{X}_{(k)} - E[\tilde{X}_{(k)}] + E[\tilde{X}_{(k)}] - x_{(k)}^o]^2 \\ &= \frac{n}{6(n+1)} + \sum_{k=1}^{n}[x_{(k)}^o - \frac{k}{n+1}]^2 \\ &\approx \frac{1}{6} + \frac{1}{6} = \frac{1}{3},\end{aligned}$$

since we expect that the second term in the last equation is about the same size of $\frac{1}{6}$ if $\{x_{(i)}^o\}_{i=1}^n$ are realized from i.i.d. $U[0,1]$.

In general, by a uniform-part $U_p[a,b]$, we denote a uniform distribution on $[a,b]$ in a possibly gapped histogram. Thus, a set of observations in a bin $[a,b]$ in a histogram will be from $U_p[a,b]$, if its decoding error is about $\frac{1}{3}(b-a)^2$. This is a definite property requirement, which we call the 'DESS criterion'. As this requirement is independent of sample size for large n, it must be satisfied by all parts of the possibly gapped histogram. It is clear that the majority of candidates in the ensemble of two-layer one-dimensional Ising model is not feasible.

The next issue is the fact that a uniform-part $U_p[a,b]$ can be divided into several uniform-parts $U_p[a_j,b_j]$ with $a_j = b_{j-1}$ for all $j = 1,\ldots,J$. Then the total DESS of the collection $\{U_p[a_j,b_j]\}^J$ is about

$$\frac{1}{3}\sum_{j=1}^{J}(b_j - a_j)^2 \left(\ll \frac{1}{3}(b-a)^2\right),$$

which can be much smaller than the DESS ($\approx \frac{1}{3}(b-a)^2$) of the original $U_p[a,b]$.

On the other hand, this collection of $\{U_p[a_j, b_j]\}^J$ incurs increasing cost of the coding length for the boundaries $\{[a_j, b_j]\}^J$. So it is necessary to balance between DESS due to stochastic randomness, and complexity of deterministic structure. Let L_0 be the relative index between the costs: decoding error and coding length. Based on the criterion of model simplicity over complexity, we need to make sure that local boundary points a_js or b_js are chosen, if the following inequality holds:

$$\frac{1}{3}(b-a)^2 > \frac{1}{3}\sum_{j=1}^{J}(b_j - a_j)^2 + (J-1)L_0;$$

or,

$$\frac{1}{3}\sum_{j\neq j'}^{J}(b_j - a_j)(b_{j'} - a_{j'}) > (J-1)L_0;$$

where L_0 is a specific index measuring how many units of decoding error is 'equal to' the coding length cost of one boundary point. That is, L_0 is determined within the domain system and is independent of sample size n. Thus, this balancing criterion being independent of n is realistic and practical, but is at odds with statistical modelling and its model selection in the statistics literature. For instance, the principle of minimum description length (MDL) is to minimize with respect to k:

$$\mathrm{MDL}(k) = -\log(P_k(x|\hat{\theta}))\pi(\hat{\theta}) + \frac{k}{2}\log n + O(k),$$

where the first two terms on the right-hand side grow with sample size n. The first term accounts for the predictive errors, while the second term accounts for the $O(1/\sqrt{n})$ precision of all global parameter estimators. And the third term is nearly constant with respect to n. Here, we would like to point out a fundamental assumption: the ensemble of candidate models is independent of sample size n. This assumption is always imposed implicitly, but hardly spelled out explicitly in the statistical literature. One serious implication of this assumption is the homogeneity of data structure that does not change as the sample size increases. This homogeneity is apparently and practically not possible to hold in histogram construction.

2.3. Possibly gapped histogram and first phase of ANOHT

Given a value of index L_0, the observed data $\{x^o_{(i)}\}_{i=1}^n$, and the fact that the exhaustive search for an optimal possibly gapped histogram within the ensemble prescribed by the two-layer one-dimensional Ising model is overwhelmingly impossible, one practically feasible computational solution is to apply the HC algorithm with the recommended complete module. This choice of the module provides the needed subdividing tendency. This tendency fits the uniform distribution well in the sense that subdividing uniform distribution reduces the total decoding errors.

The computational algorithm for a small set of potentially possibly gapped histograms is proposed based on the bifurcating feature of an HC-tree as follows. Let us define a bifurcating inter-node in an HC-tree to be active, if its parent inter-node has not been marked with a STOP sign. For practical purpose, the index L_0 is used as a threshold value.

Algorithm 1. Algorithm for possibly gapped histograms.

Step-1: Compute the DESS on the single data-range, $[x^o_{(1)}, x^o_{(n)}]$, as having only one bin. If DESS criterion is satisfied, then stop this algorithm. If DESS criterion is violated, go to the Step-2.

Step-2: Find the next highest active bifurcating inter-node and compute its DESS, say $DESS(P)$. If $DESS(P) < L_0$, or DESS criterion is satisfied, then mark this inter-node with a STOP sign. Then check if there are active inter-nodes remaining in the HC-tree. If yes, then repeat to Step-2; otherwise go to the Step-3. If $DESS(P) \geq L_0$, and if DESS criterion is violated, then repeat Step-2.

Step-3: Stop when no more active inter-nodes are left. Check the existential gap between all consecutive uniform-parts.

There are several ways to check whether a gap exists between two consecutive uniform-parts of a histogram. For example, theoretically extended boundaries can be estimated via exponential distribution of the extreme random variable. If the two estimated boundaries do not cross each other, then the two uniform-parts are separated. So there exists a gap between them. Another practical way of checking is

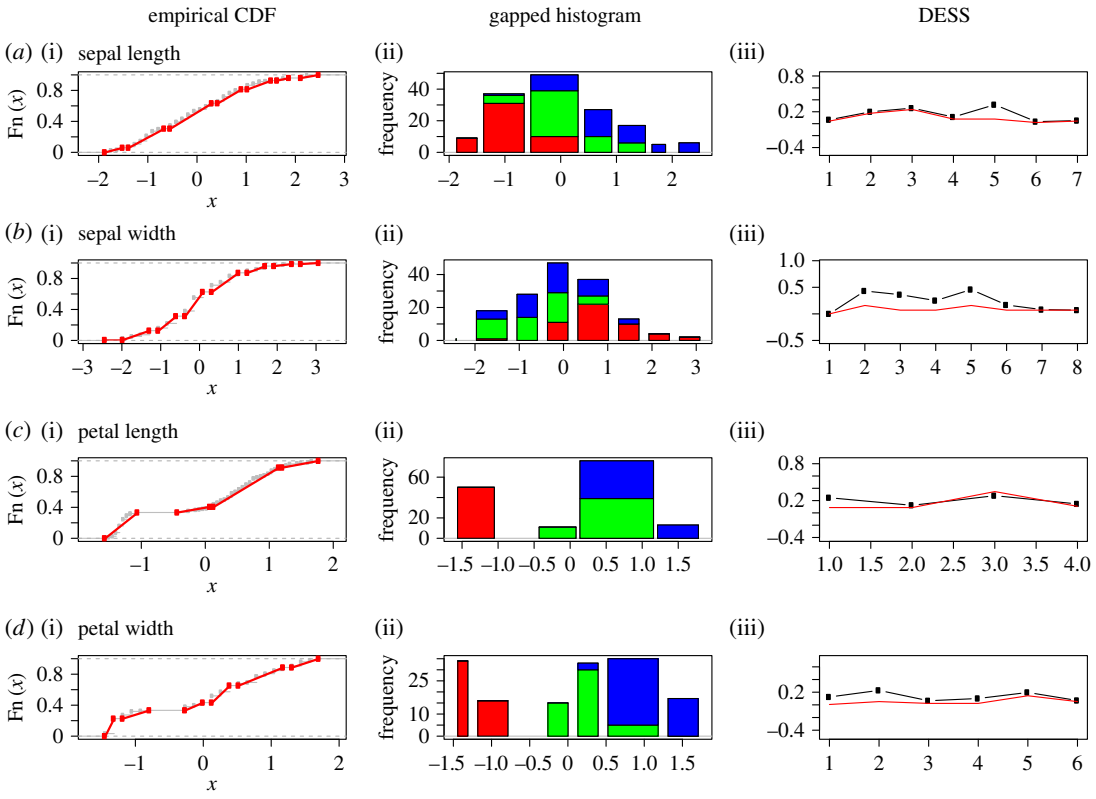

Figure 2. The gapped histograms of the four columns of *Iris* dataset (namely, petal length, petal width, sepal length and sepal width, respectively) are displayed in four rows above. The proportions of three species *setosa*, *virginica* and *versicolor* in each histogram are, respectively, represented by red, green and blue. Panels (*a*(i),*b*(i),*c*(i),*d*(i)) show the corresponding empirical cumulative distribution function (CDF), separated by red lines indicating each different bin from the histogram. Panels (*a*(ii),*b*(ii),*c*(ii),*d*(ii)) show the four possibly gapped histograms via the HC algorithm with 'complete' module and index $L_0 = 0.1 \times$ (tree height). Panels (*a*(iii),*b*(iii),*c*(iii),*d*(iii)) describe the DESS for each bin of the histograms, along with the threshold $(b_j - a_j)^2/3$ plotted in red.

to recalculate the DESS of these two uniform-parts after modifying both of them so that they share a common extended boundary, i.e. the mid-point of their extreme values. Then if both of them satisfy the DESS criterion, then there does not exist a gap. Otherwise, there exists one.

The histogram resulted from the above algorithm is the coarsest version of possibly gapped histogram. If an optimal histogram is needed, then all uniform-parts with large DESS ($>L_0$) can be bifurcated according to HC-tree branches to further improve the total decoding errors at a cost L_0 of coding one more boundary.

Furthermore, the higher an HC-tree inter-node is, the bigger is the gap potential. In fact, a gap is often visible in the coarsest version of a possibly gapped histogram, since it gives rise to one possibly gapped piece-wise linear approximation onto the empirical distribution function, see illustrations of four features of *Iris* data in figure 2. Particularly in figure 2*c*,*d* for petal length and petal width, we see evident gaps. The biological significance of the possibly gapped histograms of *Iris*' petal length and width is clearly revealed through the separation of colour-coded species.

Here, we explicitly illustrate such computations on *Iris* dataset, after standardizing each feature to zero mean and unit standard deviation. When a histogram is indeed 'gapped', this gap should be identified as a corresponding break in the empirical cumulative distribution function (CDF) also. We can verify this through the distributions of the extreme random variables of the $U(a, b)$ distribution, as the following:

$$\hat{a} = X_{(1)} - \frac{(X_{(n^*)} - X_{(1)})}{n^* + 1} \quad \text{and} \quad \hat{b} = X_{(n^*)} + \frac{(X_{(n^*)} - X_{(1)})}{n^* + 1},$$

where n^* is the bin-frequency.

As in figure 2*c*, the right boundary of the first bin of petal length can be estimated as -1.032906, while the left boundary of the second bin is estimated as -0.5127225. Also, in figure 2*d*, the right boundary of the second bin and the left boundary of the third bin of petal width are estimated as -0.7274579 and -0.3240107. The computed left and right boundary estimates do not cross each other in both cases. Hence

this implies that the gap between *setosa* and the other two species is significant. The DESS values of the final gapped bins are pretty low and fall around the uniform DESS criterion. By contrast, the *Iris*' sepal length and width do not bear such significance.

As a histogram of a feature is usually taken as a simple data-visualization step within a long and complicated process of data analysis, it hardly constitutes any stand-alone goal of data analysis. Here we would like to point out that this impression is completely incorrect. In fact, a possibly gapped histogram indeed is an important and useful tool for data analysis.

Consider a possibly gapped histogram constructed via algorithm 1 applied to a dataset by pooling measurements from J treatments. We are ready to simultaneously compare these J treatment-specific distributions by encoding J treatment-specific colours onto each bin with respect to its member-measurements' treatment identifications. Each bin has a composition of coloured proportions. So its entropy relative to the entropy of J treatment sample sizes, or their ratio is an index for bin-specific local comparison among the J treatments. A p-value can be also derived from the simulation study via simple random sampling without replacement. Hence we can afford to test these J treatments by pointing out where and how they are different. If an overall index is needed, then the weighted entropy across all bins is calculated, and so is its p-value. This is the first phase of ANOHT.

In summary, a colour-coded possibly gapped histogram can simultaneously offer more informative comparisons among J treatments than Kolmogorov–Smirnov test, which is limited for pairwise distributional comparisons, and one-way analysis of variance (ANOVA), which focuses on comparing the J mean-values.

This first phase of ANOHT on *Iris* data is shown through figure 2*a*(ii),*b*(ii),*c*(ii),*d*(ii). All four colour-coded possibly gapped histograms reveal very informative comparisons among the three *Iris* species. The compositions of bin-specific colour-codes very importantly reveal how and where these species are different. Such locality information is essential and invaluable. The biological meaning of gaps in histograms pertaining to petal length and width become evident and crucial. Here, for local testing purpose with respect to all four features, we see that majorities of p-values are either zeros or extremely small due to single colour dominance or one colour being completely missing. Likewise, the overall p-values are extremely small. Therefore, biologically speaking, these four features all provide informative pieces of information for differentiating these three species of *Iris*. This is a conclusion that has not been seen in the literature, particularly in machine learning.

2.4. Tree geometry and second phase of ANOHT

As the first phase of ANOHT provides a way of simultaneous comparison of J treatments locally and globally, the second phase of ANOHT intends to precisely address issues regarding: which treatments are closer to which and farther away from which. Our algorithm will construct a tree-geometry upon these J treatment-nodes to provide comprehensive information regarding these issues. Upon this tree geometry, this algorithm further evaluates an 'authenticity' index at every inter-node of tree-branch. Here an authentic tree branch means that its memberships are strongly supported by the data, not due to chance.

Heuristic ideas underlying this algorithm are as follows. A colour-coded possibly gapped histogram (of counts) is transformed into a matrix: one coloured bin is turned into one column according to colour-specific rows, that is, one colour for one row. Then the $J \times K$ matrix of counts is normalized row-by-row into a matrix having K columns of frequencies. Then, for simplicity, Euclidean distance is calculated between any pair of rows and the hierarchical clustering algorithm is applied to build an HC-tree upon the row axis. So a heatmap superimposed with such an HC-tree has resulted in $J-1$ inter-nodes, so $J-1$ branches.

Along the construction process of this HC-tree, each $J-1$ inter-node is associated with a tree height. We sort and digitally rank these $J-1$ tree heights from smallest to largest as a way of re-normalization. Hence each inter-node specifying a branch will be assigned with a rank-digit. This inter-node-specific rank-digit indicates that the tree height for any node within this branch to meet with any other node outside of this branch must be ranked higher.

Using the empirical process theory for complete data [8,9], mimicking is then applied on this heatmap, by simulating each row vector. The specific multivariate normality structure used here is given as follows. Let $F_n(t) = 1/n \sum_{i=1}^{n} \mathbf{1}(X_i \le t)$ be a generic empirical distribution estimating a true distribution $F(t)$ pertaining to one of the J treatments. The classical empirical process theory implies that

$$\sqrt{n}(F_n(t) - F(t)) \xrightarrow{n \to \infty} Z_F(t), \forall t,$$

where $Z_F(t)$ is a Brownian Bridge, which is a Gaussian process with covariance function $\sigma_F(t,t') = F(t)(1 - F(t'))$ when $t < t'$. Therefore, with $(t_0, t_1, t_2, \ldots, t_K)$ being the chosen ordered boundaries of K bins, the vector $(Z(t_1), \ldots Z(t_K))^{\mathrm{T}}$ comes from a family of multivariate normal distribution as below.

$$\tilde{Z}(t) = (Z(t_1), \ldots Z(t_K))^{\mathrm{T}} \sim N(0, \Sigma_K) \quad \text{with } t_1 < t_2 < \cdots < t_K,$$

where,

$$\Sigma_k = \begin{pmatrix} F(t_1)(1-F(t_1)) & F(t_1)(1-F(t_2)) & \ldots & F(t_1)(1-F(t_K)) \\ & F(t_2)(1-F(t_2)) & \ldots & F(t_2)(1-F(t_K)) \\ & & \ldots & \ldots \\ & & & F(t_K)(1-F(t_K)) \end{pmatrix}$$

$$= A\,\mathrm{diag}(F(t_1), F(t_2), \ldots, F(t_K))A^{\mathrm{T}} - \tilde{F}(t)\tilde{F}^{\mathrm{T}}(t) \tag{2.1}$$

with $\Delta_k F(t) = F(t_k) - F(t_{k-1}), k = 1, \ldots, K$. $\tilde{F}^{\mathrm{T}}(t) = (F(t_1), F(t_2), \ldots, F(t_K))$, and the matrix A and its inverse A^{-1} taking the following forms:

$$A = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix} \quad \text{and} \quad A^{-1} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{pmatrix}.$$

Thus, we have $A^{-1}\tilde{Z}(t) = (\Delta_1 Z(t), \Delta_2 Z(t), \ldots, \Delta_K Z(t))^{\mathrm{T}} = \Delta\tilde{Z}(t)$.

Further, we have the following asymptotic multivariate normality result based on equation (2.1):

$$\Delta\tilde{Z}(t) \sim N(0, \Sigma_K^*),$$

with

$$\Sigma_K^* = \mathrm{diag}(\Delta_1 F(t), \Delta_2 F(t), \ldots, \Delta_K F(t)) - \Delta\tilde{F}(t)(\Delta\tilde{F}(t))^{\mathrm{T}}.$$

This $K \times K$ covariance matrix Σ_K^* with small and negative off-diagonal entries: $-\Delta_j F(t)\Delta_{j'} F(t)$, $j \neq j'$, is the key component for mimicking each row vector of the aforementioned $J \times K$ matrix (or heatmap). An HC-tree is also derived for each mimicked heatmap. By performing such mimicking procedure many times, an ensemble of HC-trees is generated. With this ensemble, we are able to compute an authenticity index at each inter-node on the original HC-tree by counting the proportion of mimicked HC-tree having the following property: the rank-digit pertaining to the smallest branch containing all nodes belonging to the original branch being smaller or equal to the original rank-digit defining the original branch. See also the description of algorithm 2.

Algorithm 2. ANOHT.

Step-1: Build a possibly gapped histogram on the pooled data, using algorithm 1.

Step-2: Compute a matrix with the frequency of each species in the bins constructed in the gapped histogram of pooled data. For $j = 1, ..., J$ treatments and $k = 1, ..., K$ bins T_{jk} denotes the number of subjects of jth treatment falling in kth bin. $n_j = \sum_{k=1}^{K} T_{jk} =$ total frequency of jth treatment. Define $P_{jk} = \frac{T_{jk}}{n_j}, \forall j, k$. Build a tree on the row-axis of the $J \times K$ matrix T. This tree will be the point of reference.

Step-3: Mimicking T by randomly simulating a $J \times K$ matrix $M = [M_{jk}]$ with its jth row $\widetilde{M}[j, .]$ being generated with respect to

$$M[j, .] = (n_j\hat{P}_{j1}, ..., n_j\hat{P}_{jK}) = n_j\widetilde{\hat{P}}_j, \text{ where } \widetilde{\hat{P}}_j \sim \text{Multi} - \text{Normal}\left(\widetilde{P}_j, \frac{1}{n_j}\Sigma_K^*\right)$$

Step-4: Generate the matrix $M = [M_{jk}]$ 10 000 times and for each generated matrix build a tree on the rows and check if the branches of these new trees are similar to the branching of the reference tree. For each bifurcating inter-node of the reference tree, calculate an authenticity index as the percentage of times the nodes re-appeared together in the mimicked trees.

It is emphasized once more that the algorithmically computed tree geometry would allow us to see which branch of treatment-nodes is authentic in the sense that within-branch distances are smaller than

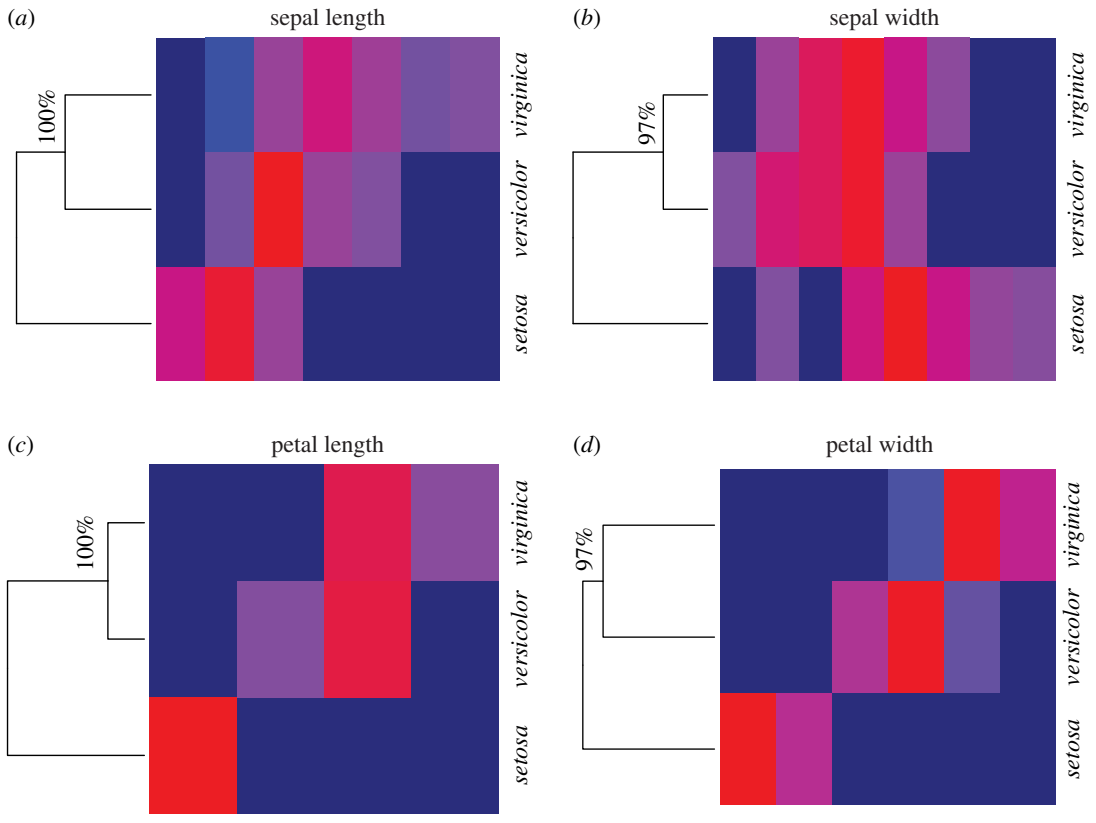

Figure 3. ANOHT for all columns of *Iris* dataset. Out of 10 000 re-generations, the nodes *versicolor* and *virginica* stay together 100% of the times in the sepal length and petal length trees (97% for and petal width trees), and in those cases *setosa* stays in a different branch. A colour closer to red represents higher value and dark blue corresponds to the lowest frequency zero.

between-branch distances with significantly high probabilities. In other words, the formation of branch-specific memberships is not likely to be due to noises. This authenticity index evaluates and confirms potential biological or mechanistic basis for this branch. This serves as one part of the knowledge discovery.

By individually applying algorithm 2 based on all four histograms in figure 2, we obtain four tree geometries on three *Iris* species-nodes with respect to their four features, as shown in figure 3. In each of the four features of *Iris*, about 97–100% of the 10 000 generations of the species tree confirm that the branch of *virginica* and *versicolor* stayed together, and the *setosa* stayed separated from them. Further, the heatmap clearly pinpoints where the significant differences are, so that the species *setosa* is rather distinct from the two other species. It is not unreasonable to think that each of these four HC-trees with authenticity indexes will shed light on the phylogenetics of these three species. This is one of the most significant merits of the second phase of ANOHT.

3. ANOHT on right-censored data

When data are compromised, such as being right-censored in survival analysis, the construction of a histogram is rarely seen in practice [10] as well in the literature [11–13]. We propose to construct a possibly gapped histogram and perform ANOHT based on Kaplan–Meier estimate of a survival function. Then the ANOHT is also performed based on the Nelson–Allen estimate of the cumulative hazard function. The latter analysis is much simpler and easier to apply than the former analysis. We apply both analyses on a heavily censored divorce dataset.

The foundation for the first phase of ANOHT is the fact that the Kaplan–Meier estimate has weights only on uncensored time points (see [14]). And the weighting scheme is termed redistribution-to-the-right (see [15]). Therefore, the first phase of ANOHT can be carried out by building a possibly gapped histogram on uncensored time points, and then applying the redistribution-to-the-right weighting scheme to adjust the weights on all bins. Nonetheless, the second phase of ANOHT would need a bit more complicated empirical process theoretical results, which are given below.

3.1. Empirical process theory on Kaplan–Meier estimate of the survival function

Given a treatment of interest, let the right-censored dataset be generically denoted as $\mathcal{X}\{(X_i, \delta_i)\}_{i=1}^n$ with $X_i = T_i \wedge C_i$, $\delta_i = \mathbf{1}_{\{T_i \le C_i\}}$ and $T_i \perp C_i$ (independent) for all $i = 1, \ldots, n$. The Kaplan–Meier estimate of the survival function $S(t) = P(T > t)$ is constructed as

$$\hat{S}_n(t) = \prod_{X_{(i)} \le t} \left(1 - \frac{\delta_{(i)}}{n - i + 1}\right),$$

where $\{X_{(i)}\}_{i=1}^n$ is the order-statistics ranking from the smallest to the largest, and $\delta_{(i)}$ is the corresponding censoring status of $X_{(i)}$.

The right-censored version of empirical process theory [12–14] states that

$$\sqrt{n}(\hat{S}_n(t) - S(t)) \xrightarrow{n\to\infty} Z_s(t), \quad \forall t \in \mathbb{R}^+,$$

where $Z_s(t)$ is a Gaussian process with characteristic covariance function

$$\sigma_s(t_1, t_2) = S(t_1)S(t_2) \int_0^{t_1 \wedge t_2} \frac{\mathrm{d}F_u(s)}{(1 - H(s))^2},$$

with $F_u(t) = P(T_i \le t, \delta_i = 1)$ and $1 - H(t) = P(X_i > t) = P(T_i > t, C_i > t)$. It is noted that $\sigma_F(t_1, t_2) = \sigma_s(t_1, t_2)$ when there is no censoring.

Likewise the K-dimensional vector $\tilde{Z}_s(t) = (Z_s(t_1), \ldots, Z_s(t_K))^{\mathrm{T}}$ is asymptotically normally distributed, that is:

$$\tilde{Z}_s(t) \sim N(0, \Sigma_K^{\#}),$$

where

$$\Sigma_K^{\#} = \mathrm{diag}(S(t_1), \ldots, S(t_K)) \cdot A \cdot \mathrm{diag}\left(\int_0^{t_1} \frac{\mathrm{d}F_u(s)}{(1 - H(s))^2}, \int_{t_1}^{t_2} \frac{\mathrm{d}F_u(s)}{(1 - H(s))^2}, \ldots\right)$$
$$\cdot A^{\mathrm{T}} \cdot \mathrm{diag}(S(t_1), \ldots, S(t_K)).$$

Hence we have the asymptotic normality of the k-dim vector $\Delta\tilde{Z}(t) \sim N(0, A^{-1}\Sigma_k^{\#}(A^{-1}))$. Throughout this section, we employ the following approximation:

$$\int_{t_{j-1}}^{t_j} \frac{\mathrm{d}F_u(s)}{(1 - H(s))^2} \approx n \sum_{t_{j-1} \le X_{(i)} \le t_j} \frac{\delta_{(i)}}{(n - i)(n - i + 1)}.$$

It is clear that the Kaplan–Meier estimate $\hat{S}_n(t)$ has jumps only on uncensored time points. Hence a construction of a possibly gapped histogram needs only to take one more step of re-adjusting the weighting, that is, a histogram for a right-censored dataset can be constructed via the following two steps:

(i) Build a histogram upon uncensored (or complete) data points;
(ii) Re-adjust the weights by applying the redistribution-to-the-right scheme for constructing the Kaplan–Meier estimate.

3.2. Empirical process theory on Nelson–Aalen estimate of cumulative hazard function

Owing to the Martingale central limit theory (see [12,13]), the Nelson–Aalen estimate of the cumulative hazard function $\Lambda(t) = -\ln S(t)$ is popularly used in survival analysis as an alternative to Kaplan–Meier estimate in statistical inferences (see [16–18]).

The Nelson–Aalen estimate of $\Lambda(t)$ is denoted as

$$\hat{\Lambda}_n(t) = \sum_{X_{(i)} \le t} \frac{\delta_{(i)}}{n - i + 1}, \quad \forall t \in \mathbb{R}^+,$$

and the Martingale central limit theorem assures that

$$\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda(t)) \xrightarrow{n\to\infty} Z_\Lambda(t),$$

where $Z_\Lambda(t)$ is a Gaussian process with independent increment property. That is, its covariance function is

$$\sigma_\Lambda(t_1, t_2) = \int_0^{t_1 \wedge t_2} \frac{\mathrm{d}F_u(s)}{(1 - H(s))^2} = \sigma(t_1 \wedge t_2).$$

Hence $Z_\Lambda(t)$ acts like a Brownian motion, so the components of the k-dimensional vector $\Delta\tilde{Z}_\Lambda(t) = (\Delta_1 Z_\Lambda(t), \ldots, \Delta_K Z_\Lambda(t))$ are indeed independently distributed, that is,

$$\Delta\tilde{Z}_\Lambda(t) \sim B(0, \Sigma_K^{**}),$$

with

$$\Sigma_K^{**} = \mathrm{diag}\left(\int_{t_0}^{t_1} \frac{\mathrm{d}F_u(s)}{(1-H(s))^2}, \int_{t_1}^{t_2} \frac{\mathrm{d}F_u(s)}{(1-H(s))^2}, \ldots \int_{t_{K-1}}^{t_K} \frac{\mathrm{d}F_u(s)}{(1-H(s))^2}\right).$$

Here Σ_K^* is the foundation for the second phase of ANOHT on complete data, while $\Sigma_K^\#$ and Σ_K^{**} are the foundations for the second phase of ANOHT on right-censored data.

4. Results on baseball data

In this section, we apply our algorithms to construct possibly gapped histograms and perform ANOHT on pitching data of three well-known pitchers in Major League Baseball (MLB). The three pitchers are Jake J. Arrieta (Chicago Cubs), Kyle C. Hendricks (Chicago Cubs) and Robert A. Dickey (Atlanta Braves). The first two pitchers with rather distinct pitching styles were keys to the 2016 World Series Champion won by Chicago Cubs, whose previous titles were in 1908 and 1907. The first pitcher is the 2015 Cy Young Award winner, the second one, who graduated from Dartmouth College, has a nickname 'The Professor', while the third one is the first knuckleball pitcher to win the Cy Young Award. This pitching dataset contains 18 732 pitches in the 2015–2016 season, including Arrieta's 6848 pitches.

The dataset was downloaded from the MLB official website provided by PITCHf/x system. This system is installed in all 30 MLB stadiums to track and digitally record the full trajectory of live baseball pitch since 2006. The data contain 21 features of each and every single pitch and batting results throughout every single game in the regular season. Here we only look at two important pitching characteristic features: start-speed and break-length. The start-speed is the detected speed of a baseball at the release point of a pitch, while break-length is the measured largest distance from the baseball's curved trajectory to the straight line linking its release point to the front of home plate. These two features are main characteristics of a pitch. They are related to each other in rather distinct ways among eight computer-classified pitch-types: (1) four-seam fastball (FF), (2) fastball cutter (FC), (3) fastball sinker (SI), (4) slider (SL), (5) changeup (CH), (6) curveball (CU), (7) knuckleball (KN) and (8) eephus (EP). The first three pitching types are in the category of fastball, while the last five are in the category of off-speed pitches. It needs to be kept in mind that a pitcher's repertoire of pitch-types is only a strict subset of these eight pitch-types.

The start-speed of Arrieta's fastball can go up to nearly 100 mph (miles per hour), while Dickey's knuckleball can be as slow as 65 mph. It is known in general that a pitch with higher start-speed tends to have smaller break-length. In fact, their relationship is relatively nonlinear. Hence, combinations of these two features are usually cleverly crafted by every professional baseball pitcher in order to effectively face batters. So it is of great interest to compare one single pitcher's as well as multiple pitchers' distinct pitch-types from these two aspects.

4.1. Data analysis on one single pitcher's pitch-types

We begin with our first phase of ANOHT on comparing pitch-types of a single pitcher from the start-speed and break-length aspects. For instance, to compare Arrieta's five pitch-types: FF, SI, SL, CH and CU, we apply algorithm 1 to construct two possibly gapped histograms of start-speed and break-length, as shown in figure 4*a* and *b*, respectively. In figure 4*a*, we see a clear gap around 84 mph. This gap bears a mechanistic difference between his curveball (CU) and the rest of four pitch-types, which have significantly higher speed. That is, this pitch-type is purposely and distinctively carried out by this pitcher. Hence this gap indicates a clear-cut implementation due to the pitcher's control capability.

This capability is even more strikingly demonstrated via the gap shown in the histogram of break-length, as shown in figure 4*b*. Again the pitch-type CU is completely separated from the rest of four types. It is also evident that pitch-types FF and SI are dominant in bins on right extreme of start-speed histogram and correspondingly dominant in bins on the left-extreme of break-length histogram.

In summary, these two colour-coded possibly gapped histograms clearly demonstrate the distinctive differences among Arrieta's five pitch-types from the two aspects. In other words, all bin-specific entropies are to be zeros or significantly small in comparison with the entropy of the distribution

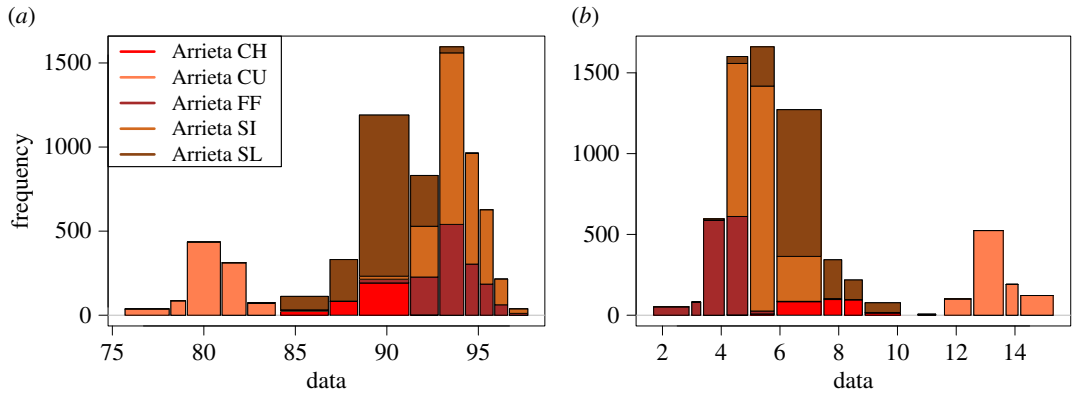

Figure 4. Possibly gapped histograms for the start-speed (*a*) and break-length (*b*) measurements of pitches thrown by Jake Arrieta in 2015–2016. The total number of pitches from three pitchers is 18 732 and out of them 6848 are from Arrieta. The five different colours in the histograms show the frequency of five different pitching styles of Arrieta in each bin.

pertaining to the five categories of pitch-types. That is, all p-values computed via the scheme of simple-random-sampling without replacement are zeros or extremely small. Similar data analysis on the other two pitchers can be likewise carried out.

4.2. Data analysis on three pitchers' pitch-types

We begin with the first phase of ANOHT to understand how similar or distinct are pitcher-specific pitch-types from the aspects of start-speed and break-length. On top of Arrieta's five pitch-types, Hendricks has five pitch-types: FF, FC, SI, CH and CU, and Dickey has three types: FF, KN and EP. So there are 13 pitcher-specific pitch-types (or treatments) in total for comparison. One coarse and one fine resolution of possibly gapped histograms of start-speed are constructed and reported in two rows of triplet panels of figure 5, respectively. After applying algorithm 1 with a choice of $L_0 = 0.1 \times$ tree height, there are 20 bins selected in this coarse version, as shown in figure 5*a–c*. The piecewise linear approximation on the empirical distribution seems reasonable, as shown in figure 5*a*, and histogram reveals evident two, or potentially three modes in figure 5*b*. But the all DESS values are relatively high in figure 5*c*.

Therefore, in order to drive DESS values lower, we select a lower L_0 and get a histogram with a fine resolution, as shown in figure 5*d–f*. Though the piecewise linear approximation becomes somehow overwhelmingly detailed with 80 bins in figure 5*d*, the histogram is seen with evident three modes in figure 5*e* and corresponding DESS values become significantly smaller than that in figure 5*c*. That is, the fine resolution histogram with very small bins on its right seems to give rise to more detailed distributional structures than the coarse one. This coarse-versus-fine resolution of histograms illustrates why we need to have data-driven bins with data-driven sizes.

Then we colour-code the coarse version of histogram, instead of the fine resolution version, for better visualization when we set to compare these 13 pitcher-specific pitch-types, as shown in figure 6*a*. Upon this histogram of start-speed, we see that bins on its left-hand side are dominated by Dickey's KN, and bins on its right-hand side are primarily dominated by Arrieta's FF and SI. While Hendrick's pitch-types are prevalent on bins situated at both sides of major valley of this histogram. It is clear that, by focusing on where differences actually occur, this colour-coded histogram is much more informative than traditional boxplot, as shown in figure 6*c*, or ANOVA, which focuses merely on comparisons of mean values.

Then we turn to the second phase of ANOHT to see what the tree geometries of the 13 pitcher-specific pitch-types look like from the aspects of start-speed and break-length. Here, it is worth re-emphasizing the significance of tree geometry. Tree geometry is a data-driven structure that makes performing branch-based-group comparison possible. A branch-based-group specified by an inter-node is confirmed as being authentic, if its branch formation is significantly supported by the data, and not by chance. This confirmation equivalently implies that, with significantly high probability, no tree nodes outside of such a group have smaller distances (or higher similarity) than distances (or similarity) among its group-member-nodes. From mechanistic perspective, such a concept of authenticity of tree branch (-based-group) has a direct implication on 'phylogenetic' information contents regarding the 13 pitcher-specific pitch-types, as would be seen below. From inference perspective, its functions have gone beyond

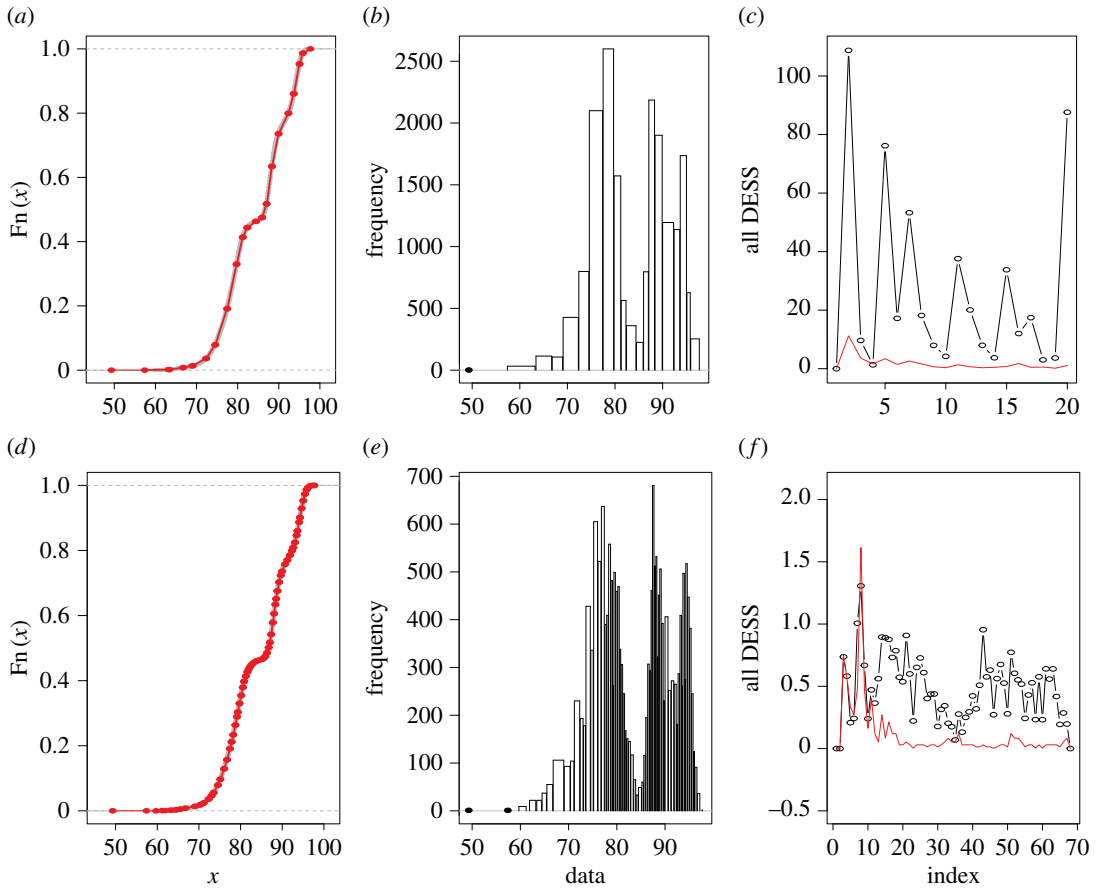

Figure 5. Possibly gapped histograms on start-speed measurements of all pitches thrown by Kyle Hendricks, Jake Arrieta and R.A. Dickey in 2015–2016. (*a–c*) A coarse version with high decoding error allowed, and (*d–f*) a finer version where decoding error is constrained to be less than 2.

multiple comparisons, such as Tukey's pairwise comparison and others, typically performed in analysis of variance (ANOVA) in classical statistics.

We apply algorithm 2 based on the coarse resolution version of histogram with 20 bins. The results based on fine resolution version of histogram with 80 bins are rather similar, but a bit harder to visualize. The resultant heatmap is framed by an HC-tree on its row axis, as reported in figure 6*b*. Percentages attached to each inter-node of the HC-tree are calculated from 10 000 mimicked HC-trees. We see via the HC-tree branches in the order going from top to bottom that the branch-based-group of Hendrick's CH and Arrieta's CU, the group of Hendrick's CU and Dickey's KN, the group of Arrieta's FF and SI, and the group of Hendrick's SI and FC, are all confirmed as being authentic with significantly high probabilities. These memberships of the four pairs of pitcher-specific pitch-types are not mixing with any pitcher's pitch-type outside of their groups. Also the triplet branch-based-groups: Hendrick's CH, Arrieta's CU and Dickey's FF and the branch-based-group of five: Hendrick's FF, SI and FC, and Arrieta's SL and CH, are also confirmed as being authentic as well. Therefore, we may conclude that these two authentic branches: one on the slow end and one on the fast end of start-speed, are two large data-driven anchors of the tree geometry.

Results from ANOHT on break-length are reported in figure 7*a*, *b*. The histogram in figure 7*a* reveals that bins on the lower end are dominated by Arrieta and Hendrick's fastball, while bins on the larger end are dominated by Arrieta's CU and Dickey's KN. Also, it is odd, but interesting to see that the bin located at 10 is nearly exclusively occupied by Dickey's KN. This exclusiveness implies that the knuckleball-specific range of break-length is probably jointly achieved by the pitcher's unusual pitching mechanics and the unusual low start-speed as seen in bins on the lower end of histogram of start-speed in figure 5*a*. This is another mechanistic pattern that can be possibly derived from ANOHT, but hardly could be easily derived from other methodologies.

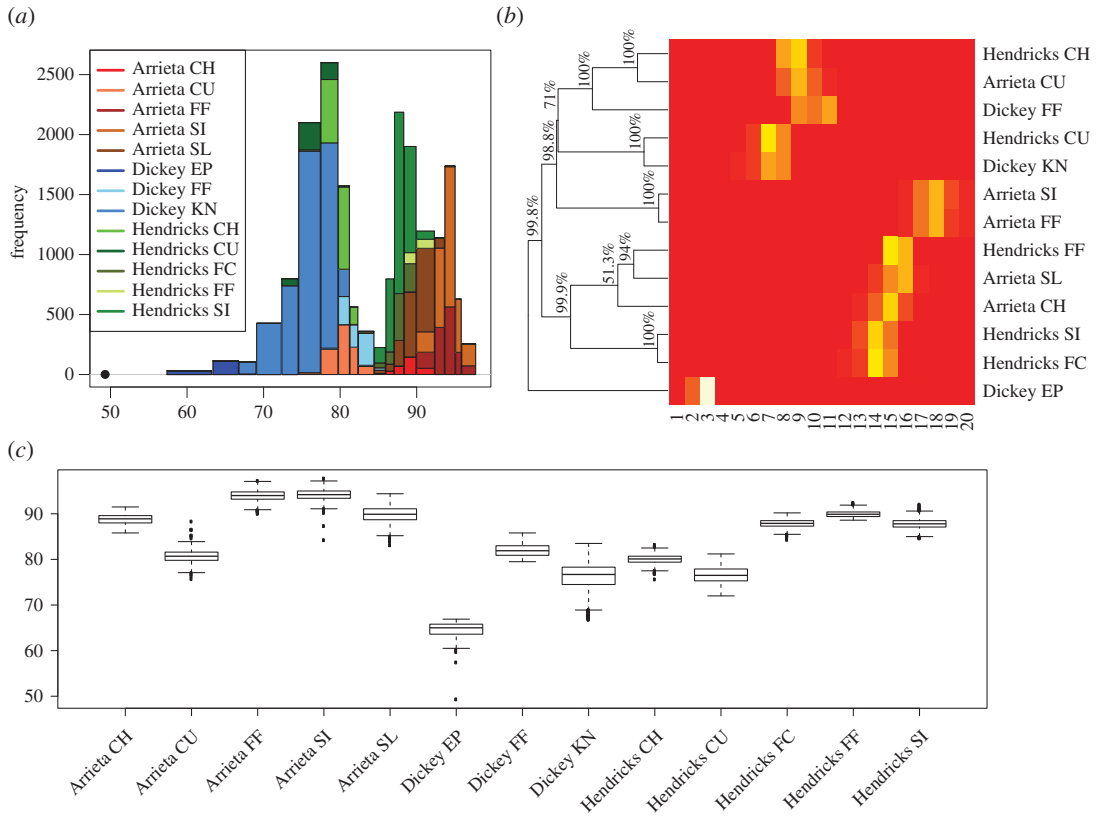

Figure 6. Coarse version of the possibly gapped histogram of start-speed from figure 5, colour-coded according to the 13 different pitches thrown by the three pitchers in (*a*). For differentiating between the three pitchers, we used shades of red, blue and green for Arrieta, Dickey and Hendricks, respectively. (*b*) The tree on ANOHT matrix with the branch strength percentages, and (*c*) shows the traditional boxplot of all start-speeds of 13 pitching styles.

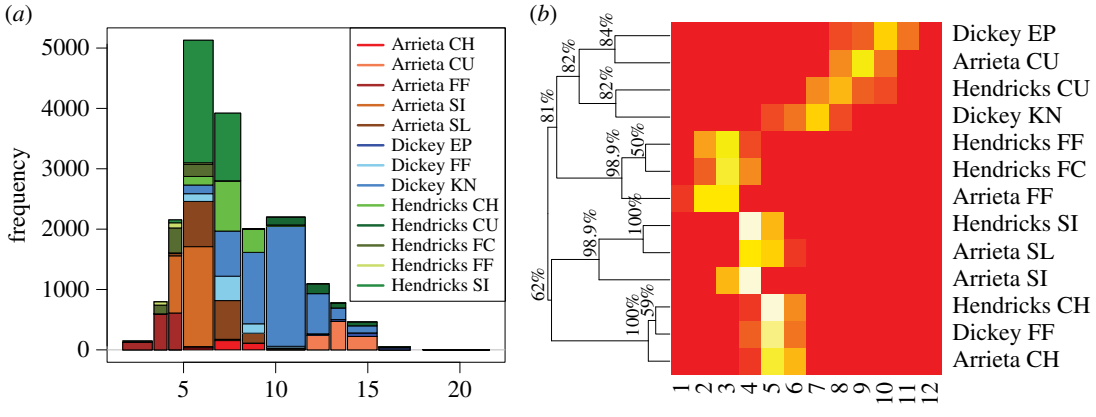

Figure 7. The possibly gapped histogram in panel (*a*) for break-length of pitches thrown by Arrieta, Dickey and Hendricks in 2015–2016, colour-coded according to the 13 different pitching styles. Panel (*b*) shows the ANOHT tree percentages on the proportion matrix of 13 pitching styles.

For tree geometry among the 13 pitcher-specific pitch types, the HC-tree superimposed on the row axis of heatmap reveals three authentic triplet branch-based-groups, as shown in figure 7*b*. These three triplet groups are: (i) Hendrick's FF and FC and Arrieta's FF; (ii) Hendrick's SI and Arrieta's SL and SI (iii) Hendrick's CH and Dickey's FF and Arrieta's CH. All these three authentic branches are all on the lower end of break-length. By contrast, indexes pertaining to the branch-based-group memberships on the higher end are high, but not high enough to be comfortably confirmed as being authentic.

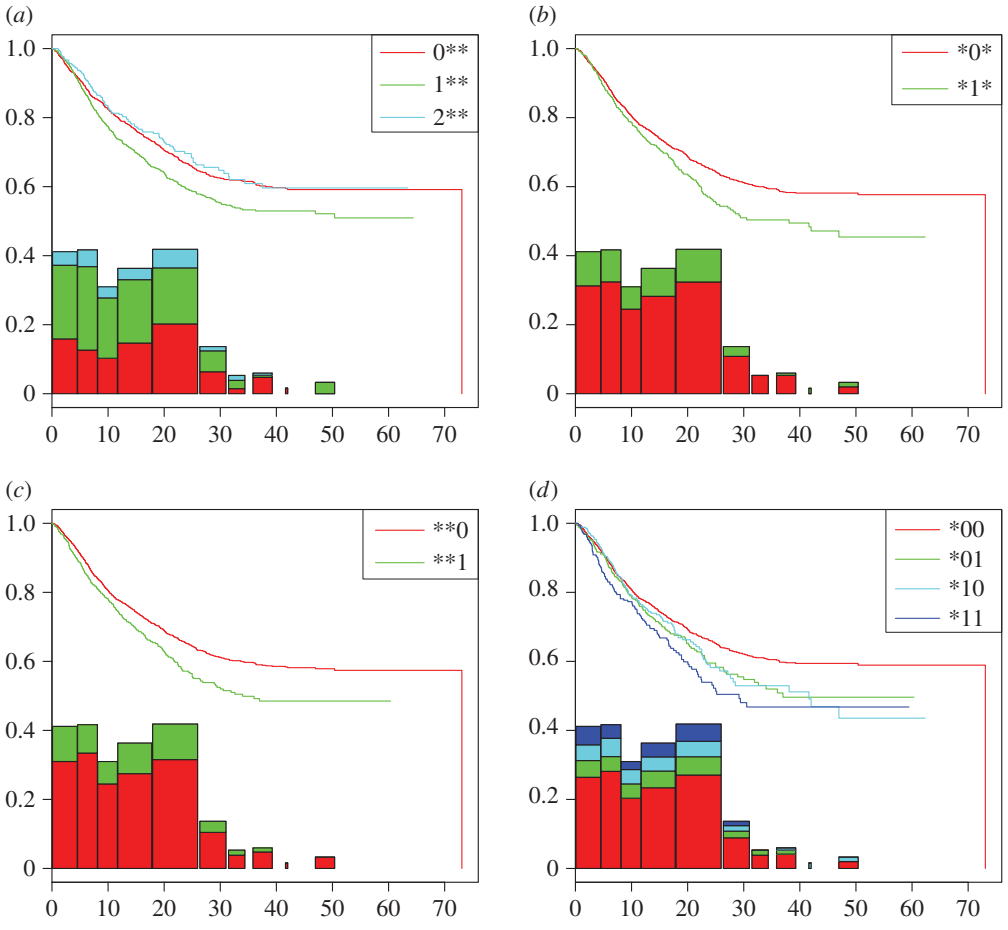

Figure 8. The first phase of ANOHT results on divorce dataset: (*a*) three treatments via first factor, (*b*) second factor, (*c*) third factor and (*d*) four treatments with respect to the second and third factors.

5. Results on divorce data

We then apply our ANOHT on a heavily censored divorce dataset. One significant feature of such a dataset is that the heavily right-censoring typically causes many issues in parametric and semi-parametric inferences in survival analysis literature [19,20]. However, it seems to have relatively little effects on ANOHT. The key reason is that ANOHT can be based only on the reliable portion of histograms.

The dissolution of marriage dataset is based on a survey conducted in the US studying 3371 couples. The unit of observation is a couple and the event of interest is divorce. Couples lost to follow-up or widowed are treated as censored observations. We have three fixed covariates: education of the husband and two indicators of the couple ethnicity regarding whether the husband is black and whether the couple is mixed.

This dataset consists of only 1033 couples with completely observed event-time of dissolution of marriage in years, and 2338 couples with right-censored event-times. The censoring rate is about $\frac{2}{3}$. The first factor is the years of husband's education with three categories: (1) coded as 0 for being less than 12 years; (2) coded as 1 for being between 12 and 15 years; (3) coded as 2 for being 16 or more years. The second factor is husband's ethnicity with two categories: (1) coded as 1 for the husband being black; (2) codes as 0 otherwise. The third factor is couple's ethnicity with two categories: (1) coded as 1 if the husband and wife have different ethnicity; (2) coded 0 otherwise. Censoring status as usual is coded as 1 for divorce and 0 for censoring (due to widowhood or lost to follow-up).

Thus, this dataset can be partitioned with respect to one single factor, or a pair of factors, or the three factors together. That is, the finest partition has 12 samples or treatments, in which each treatment is triple-coded. For instance, the treatment '201' stands for the sample of couples with husband's education

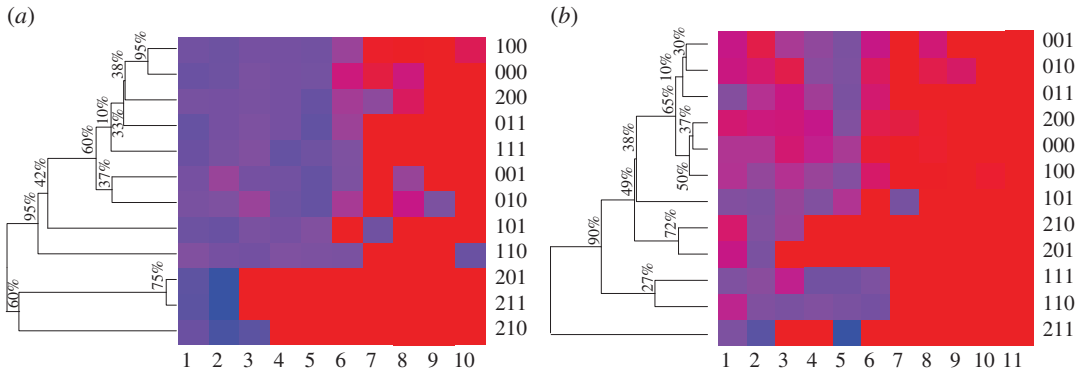

Figure 9. The second phase of ANOHT results on divorce dataset: (*a*) based on Kaplan–Meier survival function estimates on 12 categories and (*b*) based on the Nelson–Aalen cumulative hazard function estimates.

being more than 15 years (coded 2 in the first factor), non-black husband (coded 0 in the second factor) and having mixed ethnicity (coded 1 in the third factor).

5.1. ANOHT on a divorce data

The first phase of ANOHT on this divorce dataset on these four aspects are reported in figure 8. Overall the possibly gapped histogram shows three evident gaps along its right tail. These gaps are not likely to be due to old age. As they are followed by bins with significant heights, their locations seem interesting, but need more research attentions for pertinent interpretations. Specifically, figure 8*a* shows three Kaplan–Meier estimates of three survival functions with respect to three categories of education levels. It is rather interesting to note that the marriage with husbands having middle education level (1**) is likely to fail much earlier than marriages with husbands in either the lowest (0**) or highest (2**) education levels. Figure 8*b*,*c* clearly reveals that the two categories: Black-husband (*1*) and mixed-ethnicity of couple (**1), respectively, associate with lower survival rates. Further the category: black-husband and mixed-ethnicity of couple (*11) have the lowest survival rate against the other three categories in figure 8*d*. This survival rate is especially lower than the one belonging to the category: non-black husband and wife (*00).

The second phase of ANOHT on divorce dataset is reported in figure 8:

The group of six treatments (categories): (000, 001, 010, 100, 011, 200) are shown in figure 9*a*,*b*. They have similar Kaplan–Meier survival function and Nelson–Aalen cumulative hazard functions. Particularly, its subgroup of three treatments: (000, 100, 200) shows very high similarity on both aspects among them. Such evidence is interpreted as that the factor of education has no effect on occurrence of event of marriage failure in non-black couple. By contrast, the presence of the subgroup of three treatments: (010, 001, 011) seems to imply that the occurrence of the event of marriage failure for a couple with husband having lowest education level is neither affected by the husband's ethnicity, nor the couple's mixture of ethnicity.

From Kaplan–Meier survival function perspective, the tree on the heatmap of figure 9*a* evidently shows one 95% authenticity of the large branch of 9 treatments against the other branch of three treatments: (201, 211, 210). Another 95% authenticity is found on the branch of two treatments: (100, 000). By contrast, from Nelson–Aalen cumulative hazard function perspective, there is only 90% authenticity found on the outlier treatment: (211) against the rest of treatments.

6. Conclusion

In this note, we demonstrate algorithmic data-driven computing for constructing a possibly gapped histogram, and then develop two phases of ANOHT. Their practical values are clearly illustrated and manifested through *Iris* data, baseball pitching data and divorce data. Our data-driven computing simply and critically demonstrates interfaces of machine learning, information theory and statistical physics. We believe such data analytic techniques would have very high potentials for very wide spectra of problems in sciences.

It is surprising that a well-known HC algorithm, which was first developed in the 1950s for taxonomy [21], can help resolve a complex computational physical problem with exponential growth of complexity. Even more striking is that such a multi-scale physical problem explicitly contrasts and brings out the unspoken assumption of homogeneity implicitly assumed in all known statistical model selection techniques.

We believe that the existential issue of gap in a distribution will become more critical in this Big Data era than ever before. Its biological and mechanistic meanings should be a part of critical knowledge discovery in data-driven analysis. Particularly, a possibly gapped histogram can provide a fundamental new method of re-normalization of data via digital coding schemes. Such a data re-normalizing method in fact plays a critical role in unsupervised machine learning for knowledge discovery.

The merits of two phases of ANOHT are apparently seen in the three real examples, so are likely to be seen in many scientific researches. We have demonstrated that they have the capabilities to resolve issues never being discussed before, and at the same time to provide knowledge discoveries from wider perspectives. Further, since they are free from common, but unrealistic constraints and assumptions imposed by statistical modelling and analysis, their results should be more authentic and closer to reality.

Here, we devote the rest of this section for implications of our developments on statistics. It is because of lacking continuity or smoothness assumptions in our developments, many types of potential complexity can be realistically embraced within our data analysis on one-dimensional dataset. For instance, the multi-scale issue is clearly seen in the histograms of start-speed, as shown in figure 4. This issue has to be resolved in a data-driven fashion because of the lack of prior knowledge. In sharp contrast, a histogram with pre-fixed number of equal-size bins is just not realistic. Here we emphasize that ignoring realistic information contents contained in data is hardly a right way of analysing real data.

This multi-scale feature also spells out the fact that all bins' boundaries are not global parameters. Thus, the ensemble of candidate histograms is far from being fixed, but grows exponentially like the ensemble of all possible spin-configurations of Ising model in statistical physics. They render no uniform $1/\sqrt{n}$ or $1/n$ rates of convergence. Thus, all statistical model selection techniques, such as AIC, BIC and minimum description length (MDL), in statistics are not applicable. The simple underlying technical reason is that all these model selection techniques require their ensemble of candidate models to be fixed and independent of n, so that the corresponding mathematical optimization problem can be well defined.

The implications of ANOHT on statistics are clearly demonstrated and contrasted through its applications on the *Iris*, MLB pitching and divorce datasets. The computing simplicity and informative results pertaining to the two phases of analyses via ANOHT come from the recognized interface of a physical problem and an unsupervised machine learning algorithm.

Data accessibility. The *Iris* data are available at UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.html. The pitching data are available in PITCHf/x database belonging to Major League Baseball via http://gd2.mlb.com/components/game/mlb/ and the subset used are available from the Dryad Digital Repository (https://doi.org/10.5061/dryad.bs632) [22]. The divorce data are available at: http://data.princeton.edu/wws509/datasets/#divorce.

Authors' contributions. H.F. designed the study. T.R. collected all data for analysis. H.F. and T.R analysed the data, interpreted the results and wrote the manuscript. All authors gave final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. We received no funding for this study.

Acknowledgements. We thank Kevin Fujii for helps in pitching data retrieval.

References

1. Cover TM, Thomas JA. 2006 *Elements of information theory*. Wiley series in telecommunications and signal processing. New York, NY: Wiley-Interscience.
2. Hall P, Hannan EJ. 1988 On stochastic complexity and nonparametric density estimation. *Biometrika* **75**, 705–714. (doi:10.1093/biomet/75.4.705)
3. Rissanen J. 1989 *Stochastic complexity in statistical inquiry*. River Edge, NJ: World Scientific Publishing Co. Inc.
4. Yu B, Speed TP. 1992 Data compression and histograms. *Probab. Theory. Relat. Fields* **92**, 195–229. (doi:10.1007/BF01194921)
5. Kontkanen P, Myllymäki P. 2007 MDL histogram density estimation. In *Proc. 11th Int. Conf. on Artificial Intelligence and Statistics* (*AISTATS-07*), *San Juan, Puerto Rico* (eds M Meila, X Shen), pp. 219–226. See http://proceedings.mlr.press.
6. Li M, Vitnyi PM. 2008 *An introduction to Kolmogorov complexity and its applications*. Berlin, Germany: Springer Publishing Co. Inc.
7. Ising E. 1925 Beitrag zur theorie des ferromagnetismus. *Z. Phys.* **31**, 253–258. (doi:10.1007/BF02980577)
8. Hsieh F. 1995 The empirical process approach for semiparametric two-sample models with heterogeneous treatment effect. *J. R. Stat. Soc. Ser. B* **57**, 735–748.
9. Shorack GR, Wellner JA. 1986 *Empirical processes with applications in statistics*. New York, NY: Wiley.
10. Kleinbaum DG. 1996 *Survival analysis: a self-learning text*. New York, NY: Springer-Verlag.
11. Cox DR. 1984 *Analysis of survival data*. London, UK: Chapman and Hall.

12. Andersen PK, Borgan O, Gill RD, Keiding N. 1992 *Statistical models based on counting process*. New York, NY: Springer.
13. Kalbfleisch JD, Prentice RL. 2002 *The statistical analysis of failure time data*. New York, NY: John Wiley.
14. Miller RG. 1981 *Survival analysis*. New York, NY: John Wiley.
15. Efron B. 1967 The two sample problem with censored data. In *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability*, vol. IV (eds LM Le Cam, J Neyman), pp. 831–853. University of California Press.
16. Hsieh F. 1996 The empirical process approach for a generalized hazards model. *Biometrika* **83**, 519–528. (doi:10.1093/biomet/83.3.519)
17. Hsieh F. 1996 Empirical process approach in a two-sample location-scale model with censored data. *Ann. Stat.* **24**, 2705–2719. (doi:10.1214/aos/1032181176)
18. Hsieh F. 2001 On heteroscedastic hazards regression model: theory and applications. *J. R. Stat. Soc. Ser. B.* **63**, 63–79. (doi:10.1111/1467-9868.00276)
19. Padgett W, Wei LJ. 1982 Estimation of the ratio of scale parameters in the two sample problem with arbitrary right censorship. *Biometrika* **69**, 252–256. (doi:10.1093/biomet/69.1.252)
20. Bassiakos YC, Meng X-L, Lo S-H. 1991 A general estimator of the treatment effect when the data are heavily censored. *Biometrika* **78**, 741–748. (doi:10.1093/biomet/78.4.741)
21. Sneath PHA, Sokal RR. 1973 *Numerical taxonomy: the principles and practices of numerical classification*. San Francisco, CA: Freeman.
22. Fushing H, Roy T. 2018 Complexity of possibly gapped histogram and analysis of histogram. Dryad Digital Repository. (doi:10.5061/dryad.bs632)