APPLIED MATHEMATICS

Semblance: An empirical similarity kernel on probability spaces

Divyansh Agarwal[1,2]* and Nancy R. Zhang[1,2]*

In data science, determining proximity between observations is critical to many downstream analyses such as clustering, classification, and prediction. However, when the data's underlying probability distribution is unclear, the function used to compute similarity between data points is often arbitrarily chosen. Here, we present a novel definition of proximity, Semblance, that uses the empirical distribution of a feature to inform the pair-wise similarity between observations. The advantage of Semblance lies in its distribution-free formulation and its ability to place greater emphasis on proximity between observation pairs that fall at the outskirts of the data distribution, as opposed to those toward the center. Semblance is a valid Mercer kernel, allowing its principled use in kernel-based learning algorithms, and for any data modality. We demonstrate its consistently improved performance against conventional methods through simulations and real case studies from diverse applications in single-cell transcriptomics, image reconstruction, and financial forecasting.

INTRODUCTION

In modern data analysis, the data are often first reduced to a proximity matrix representing the pairwise similarity between observations, which becomes the input to downstream analyses such as clustering, classification, and prediction. This proximity matrix is an information map as well as an information bottleneck—the former, because all of the information available to a downstream analysis algorithm is represented in the matrix, and the latter, because the matrix must transmit enough information about the data for any downstream method to be able to do its task. In exploratory data analysis settings, Euclidean distance–or correlation-based metrics are popular ad hoc choices for inferring proximity (*1*–*3*), although more sophisticated, context-specific choices have been designed for particular tasks (*4*, *5*).

During the past two decades, efficient kernel-based learning algorithms and their reproducing kernel Hilbert space (RKHS) interpretations have generated intense renewed interest. Specifically, efforts have focused on the development of proximity matrices that satisfy the Mercer condition, which would allow the detection of complex nonlinear patterns in the data using well-understood linear algorithms (*6*–*8*). Such proximity matrices, called Mercer kernels, form the core of several state-of-the-art machine learning systems. Constructing a similarity function or a proximity matrix amounts to encoding our prior expectation about the possible patterns we may be expected to learn in a given feature space, and thus, it is a critical step in real-world data modeling (*9*, *10*). Noise distributions in the real world are often non-elliptical, with continuous and discrete features generally intermixed. Yet, in the initial stages of data analysis, when the underlying structure of the data's probability space is unclear, the choice of the similarity/distance metric is often arbitrary. In exploratory data analysis, there is often little prior knowledge to guide the selection of distance/similarity measures, much less the design of valid Mercer kernels. Thus, even as kernel-based machine learning algorithms become sophisticated, due to the lack of more informed options, we often default to relying on Euclidean distance or Pearson correlation during the exploratory stage of data analysis.

Here, we present a general, off-the-shelf kernel function, Semblance, that uses the empirical distribution of a feature across all observations to inform the proximity between each pair. Semblance puts a premium on agreement in rare feature values, for discrete features, and on proximity between points at the outskirts of the data distribution, for continuous features. This allows Semblance to reveal structures in the data that are obscured by current, commonly used kernel measures. We first describe the intuitions behind Semblance using a concrete example and subsequently prove that it is a valid Mercer kernel and thus can be used in any kernel-based learning algorithm (*11*). Then, under simplified but transparent simulation experiments, we systematically explore the types of patterns that we can expect to identify using Semblance versus other common approaches such as Euclidean distance. Semblance achieves higher sensitivity for niche features by adapting to the empirical data distribution. Through examples from several fields—single-cell biology, image analysis, and finance—we demonstrate how the Semblance kernel can be used.

Constructing the rank-based semblance function

Suppose we begin with $N_{n \times G}$, the data matrix with n rows and G columns. Let each row correspond to an object and each column correspond to a feature measured for all objects. We would like to construct a similarity kernel relating the objects. For ease of notation, let $X = (x_1, \ldots, x_g, \ldots, x_G)$ and $Y = (y_1, \ldots, y_g, \ldots, y_G)$ be two objects, i.e., two rows in the matrix $N_{n \times G}$.

Consider a feature for which most objects record the value "1," and only very few record the value "0." Now, consider two objects, both of which record the rare value "0" for this feature. Is this stronger evidence for similarity between these two objects, as opposed to the scenario where both record the much more typical value "1"? Intuitively, it is much more improbable for two independent objects to agree on a rare value than on a common value, and thus, two objects agreeing on the rare values for many features suggests that they may belong to a common niche group. Similarly, for a continuous-valued feature, proximity in terms of absolute distance between two independent draws is much more unlikely at the tails of the distribution as compared to at its center. Thus, for discrete features, we would like to reward agreement

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA. [2]Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
*Corresponding author. Email: divyansh.agarwal@pennmedicine.upenn.edu (D.A.); nzh@wharton.upenn.edu (N.R.Z.)

between objects on rare feature values, and for continuous features, we would like to reward proximity between objects in the tail of the empirical distribution. Furthermore, for robustness, it is desirable for the similarity function to be nonparametric and invariant to monotone transformations of the features. These are the considerations underlying the construction of Semblance.

More formally, for a given feature g, let $\mathbb{P}_g$ be its underlying probability distribution. Let the observed values for this feature g in objects X and Y be x_g and y_g, respectively. In practice, we do not know $\mathbb{P}_g$, but if we did, we could ask how likely are we, if we were to redraw one of the two values (x_g, y_g), to obtain a distance that is equal to or smaller than the actual observed distance, while preserving the order between the two. Let Z be the redraw, then this could be expressed as the probability

$$p_g(x_g, y_g) = p_g(y_g, x_g) := \mathbb{P}_g\{\min(x_g, y_g) \leq Z \leq \max(x_g, y_g)\} \quad (1)$$

The above probability is a natural measure of the dissimilarity between any two values of feature g (see Fig. 1). A subtle but important detail is that the probability in Eq. 1 includes both endpoints x_g and y_g, and therefore $p_g(x_g, x_g) = \mathbb{P}\{x_g\} > 0$. This definition of proximity is desirable because it naturally incorporates the information in the underlying probability measure that generated the data. For example, as illustrated in Fig. 1, in the binary setting, it is much more rare for two observations to both be equal to 0 if 0 has low probability, and thus, the "reward" for $x_g = y_g = 0$ depends on the probability mass at 0. Similarly, in the continuous setting, the reward for proximity between x_g and y_g depends on where the pair falls on the distribution. For the same linear distance between x_g and y_g, their dissimilarity is higher when they both fall at the center of the distribution than when they both fall at the tails.

In practice, $\mathbb{P}_g$ is not known, but with a large enough sample size, the empirical distribution $\hat{\mathbb{P}}_g$ serves as a good approximation, leading to the plug-in empirical estimate $\hat{p}_g(x_g, y_g)$, obtained by substituting $\hat{\mathbb{P}}_g$ for $\mathbb{P}_g$ in Eq. 1. This is reminiscent of empirical Bayes methods, where information is borrowed across all observed values to inform our dissimilarity evaluation between any given pair. We define

$$k_g(X, Y) = 1 - \hat{p}_g(x_g, y_g) = \frac{1}{n}\sum_{i=1}^{n}[1 - I(\min(x_g, y_g) \leq N_{ig} \leq \max(x_g, y_g))]$$

the empirical probability of falling strictly outside the interval $[x_g, y_g]$. The indicator I returns 1 if $\min(x_g, y_g) \leq N_{ig} \leq \max(x_g, y_g)$, and 0 otherwise.

Suppose feature g is continuous, and hence each observed value is unique, and let r_X, r_Y be the ranks of x_g, y_g among all observed values of this feature across the n objects. Then, k_g can be expressed simply as a function of the ranks

$$k_g(X, Y) = \frac{1}{n}(|r_X - r_Y| + 1)$$

However, for discrete features, the computation of $k_g(X, Y)$ is more complicated owing to ties. Nevertheless, computing $k_g(X, Y)$ in general is easy and fast. An example algorithm is provided in Materials and Methods.

We now define the Semblance function as

$$K(X, Y) = \frac{1}{G}\sum_{g=1}^{G} k_g(X, Y) w_g \quad (2)$$

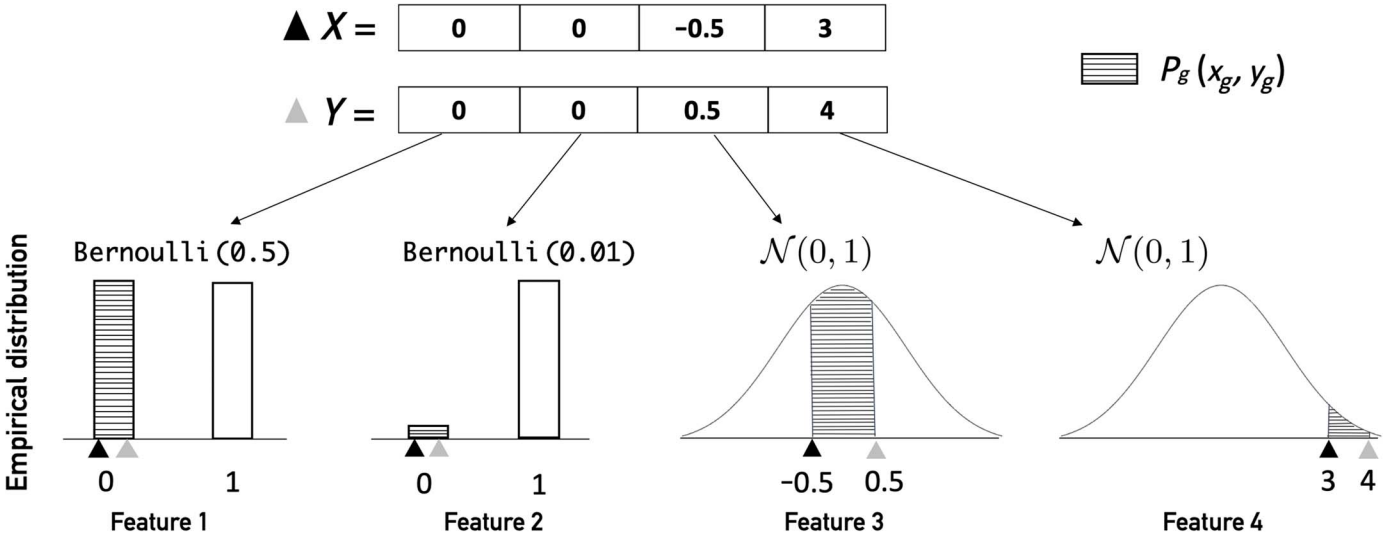


Fig. 1. Illustration of what $p_g(x_g, y_g)$ corresponds to in the case of a discrete distribution or a continuous distribution. In this toy example, X and Y are two objects with four features measured. Semblance computes an empirical distribution from the data for each feature and uses the information of where the observations fall on that distribution to determine how similar they are to each other. Specifically, it emphasizes relationships that are less likely to occur by chance and that lie at the tail ends of a probability distribution. For example, X and Y are equal to 0 for both the first and second feature, but these two features contribute different values to the kernel: "0" is more rare for the second feature, and thus p_2 (0, 0) is smaller than p_1 (0, 0), and the second feature contributes a higher value in the Semblance kernel. Similarly, even though the difference between X and Y is 1 for both features 3 and 4, feature 4, where the values fall in the tail, has lower $p_g(x_g, y_g)$ and thus contributes a higher value in the Semblance kernel than feature 3.

where w_g corresponds to the relative weight or importance of each feature. When there is reliable domain knowledge to prioritize features, these should be used to construct the weights. When no a priori information is available, a weight that reflects the shape of the feature distribution can be used, for example, the Gini coefficient for positive-valued features or a robust approximation to the negentropy for real-valued features. In our experiments, we have found that considering all features to be equally important ($w_g = 1$) gives decent results in most cases.

Since $\hat{p}_g(x_g, x_g) \leq \hat{p}_g(x_g, y_g)$ if $x_g \neq y_g$, it follows that $K(X, X) \geq K(X, Y)\ \forall\ X \neq Y$.

Thus, when applied to any data matrix N, this function outputs a symmetric $n \times n$ matrix whose rows and columns are maximized at the diagonal.

RESULTS

Semblance is a valid Mercer kernel

Since $K(X, Y)$ is just the mean of $k_g(X, Y)$ across g, we start by considering

$$K_g = \{K_g(i,j) = k_g(N_{ig}, N_{jg}) : 1 \leq i, j \leq n\}$$

the matrix derived only from observations of feature g. First, assume that the objects have been permuted such that $\{N_{ig} : i = 1, \ldots, n\}$ are monotone nondecreasing. Define

$$a_i = \hat{\mathbb{P}}(Z < N_{ig}) \text{ and } b_i = \hat{\mathbb{P}}(Z > N_{jg}) \tag{3}$$

suppressing the notational dependence of a_i and b_i on g, for simplicity. On the basis of Eq. 1, for $i \leq j$

$$K_g(i,j) = a_i + b_i \tag{4}$$

By our monotone nondecreasing assumption, $a_i \leq a_{i+1}$ and $b_i \geq b_{i+1}$. Thus, K_g has the decomposition

$$K_g = \begin{bmatrix} a_1 & a_1 & \ldots & a_1 \\ a_1 & a_2 & \ldots & a_2 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_1 & a_2 & \ldots & a_n \end{bmatrix} + \begin{bmatrix} b_1 & b_2 & \ldots & b_n \\ b_2 & b_2 & \ldots & b_n \\ b_3 & b_3 & \ldots & b_n \\ \vdots & & \ddots & \vdots \\ b_n & b_n & \ldots & b_n \end{bmatrix} = M + N$$

Remark: The matrices M and N have a symmetric and analogous structure. The left upper hook comprising the first row and column of M has all entries a_1, the second hook has all entries a_2 and so on, until the nth hook, which is simply the entry a_n. Similarly, the right lower hook of N comprising the last row and column has all entries b_n, all the way up to the solo entry b_1 in the first row and column. *Proposition 1*: M is a nonnegative-definite (NND) matrix.

The proof is by induction. For the base case, consider the 2×2 matrix

$$\dot{M} = \begin{bmatrix} a_1 & a_1 \\ a_1 & a_2 \end{bmatrix}$$

By construction $a_{i+1} \geq a_i$, therefore $det(\dot{M}) \geq 0$, and hence $\dot{M}$ is NND. The induction hypothesis is that all $m \times m$ matrices, $\bar{M}$, with the structure

$$\begin{bmatrix} a_2 & a_2 & \ldots & a_2 \\ a_2 & a_3 & \ldots & a_3 \\ a_2 & a_3 & \ldots & a_4 \\ \vdots & & \ddots & \vdots \\ a_2 & a_3 & \ldots & a_m \end{bmatrix} \text{ where } a_2 \leq a_3 \leq \ldots \leq a_m \tag{5}$$

are NND. Now, to prove that the same is true for $m \times m$ matrices, we can write such matrices in the form

$$\begin{bmatrix} a_1 & U \\ U^T & \bar{M} \end{bmatrix} \tag{6}$$

where U represents the vector $(a_1 a_1 a_1 \cdots)$ and $\bar{M}$ is a matrix that satisfies the induction hypothesis. Using the Schur complement condition for the nonnegative definiteness of a symmetric matrix (*12*), we can show that $\bar{M} - U^T {a_1}^{-1} U$ is NND

$$\bar{M} - U^T {a_1}^{-1} U = \bar{M} - \begin{pmatrix} a_1 \\ a_1 \vdots \\ a_1 \end{pmatrix} \frac{1}{a_1} (a_1 a_1 a_1 \cdots)$$

$$= \bar{M} - \begin{bmatrix} a_1 & a_1 & \ldots & a_1 \\ a_1 & a_1 & \ldots & a_1 \\ \vdots & & \ddots & \vdots \\ a_1 & a_1 & \ldots & a_1 \end{bmatrix}$$

$$= \begin{bmatrix} a_2 - a_1 & a_2 - a_1 & \ldots & a_2 - a_1 \\ a_2 - a_1 & a_3 - a_1 & \ldots & a_3 - a_1 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_2 - a_1 & a_3 - a_1 & \ldots & a_n - a_1 \end{bmatrix}$$

This resultant matrix is of a form that satisfies Eq. 5 and thus, by the induction hypothesis, is NND. Therefore, the matrix (Eq. 6) is also NND.

Since N mirrors the properties of M by construction, we have by *Proposition 1* that N is also an NND matrix. For NND matrices, it is also true that (i) the sum of NND matrices is NND, and (ii) applying the same permutation to the rows and columns of an NND matrix preserves the NND structure (see proofs in Materials and Methods). On the basis of these facts, together with *Proposition 1*, the kernel matrix K (sum of all K_g's) is NND. The matrix K computed on any data matrix by the Semblance function defined in Eq. 2 is NND, and thus, Semblance is a valid Mercer kernel. As a result, the Representer theorem allows effective implementation of nonlinear mappings through inner products represented by our kernel function (*6*, *11*). We review the theory governing the existence of an RKHS and a feature space for Semblance in the Supplementary Materials.

Semblance is conceptually different from rank-based similarity measures

Since, in the case where all features are continuous, Semblance can be simplified to a function on ranks, we first clarify how it differs from existing rank-based similarity measures: Spearman's rho (ρ) and

Kendall's tau (τ). By construction, Semblance is fundamentally different from these existing measures in two ways. First, while ρ and τ are based on ranks computed by ordering the values within each object (the rows of matrix N), Semblance is computed using ranks determined by ordering the values within each feature (the columns of matrix N). Thus, the Semblance kernel can be expected to produce values that differ substantially from these two measures. Second, Semblance treats ties differently from simple rank-based methods, such that ties shared by many objects diminish the proximity between those objects. This treatment of ties, for discrete data, makes Semblance more sensitive for niche subgroups in the data. Therefore, Semblance is better understood through the lens of empirical Bayes, where, for each feature, the empirical distribution across all objects informs our evaluation of the similarity between each pair of objects.

Simulations

Simulations allow us to compare the effectiveness of similarity/distance measures under simplified but interpretable settings. We used simulations to compare Semblance against Euclidean distance, Pearson correlation, and Spearman correlation in their ability to separate two groups in an unsupervised setting. We simulated from a two-group model, where multivariate objects came from either group 1, with probability $q < 0.5$, or group 2, with probability $1 - q$. Let each object contain m features, drawn independently, with a proportion $p \in (0, 1)$ of the features being informative. The informative features have distribution $P_{I,1}$ in group 1 and $P_{I,2}$ in group 2. The rest of the features are noninformative and have the same distribution P_{NI} across both groups. We consider both continuous and discrete distributions for the features. In the continuous case, the features are generated from

$$P_{NI} = N(0,1), \quad P_{I,1} = N(\mu\sigma_2, \sigma_1), \quad P_{I,2} = N(0, \sigma_2) \qquad (7)$$

In the discrete case, the features are generated from

$$P_{NI} = P_{I,2} = \text{Bernoulli}(r_0), \quad P_{I,1} = \text{Bernoulli}(r_1) \qquad (8)$$

Of course, whether a feature is informative or not, and whether an object is from group 1 or group 2, is not used when computing the similarity/distance matrix.

As shown in Fig. 2A, in each simulation run, we generated n objects with the first $n_1 = qn$ coming from group 1 and the next $n_2 = (1 - q)n$ coming from group 2. Our goal is to detect the existence of the minority group 1 and assign objects to the appropriate group. Similarities (Semblance, Pearson, and Spearman) and distances (Euclidean) are computed on these data, each producing an $n \times n$ matrix, which we will call S. Let

$$\bar{S}_{11} = \frac{1}{n_1} \sum_{1 \le i < j \le n_1} S_{ij}, \quad \bar{S}_{22} = \frac{1}{n_2} \sum_{n_1 < i < j \le n} S_{ij},$$
$$\bar{S}_{12} = \frac{1}{n_1 n_2} \sum_{1 \le i \le n_1 < j \le n_2} S_{ij}$$

Then, $\bar{S}_{11}$ is the mean similarity/distance between objects in group 1, $\bar{S}_{22}$ is the mean similarity/distance between objects in group 2, and $\bar{S}_{12}$ is the mean similarity/distance across groups. To quantify the signal in S, we let $T_1 = (\bar{S}_{11} - \bar{S}_{12})/\text{se}_1$, $T_2 = (\bar{S}_{22} - \bar{S}_{12})/\text{se}_2$, where se_1 and se_2 are standard errors of the differences in the numerators. Hence, large positive values of T_1 and T_2 imply that downstream algorithms based on S will be able to separate the two groups well.

Figure 2B shows the T_1 and T_2 values for an example set of simulations where $n = m = 100$, the proportion of informative features is 10%, the rare subpopulation proportion is 10%, and every feature is normal

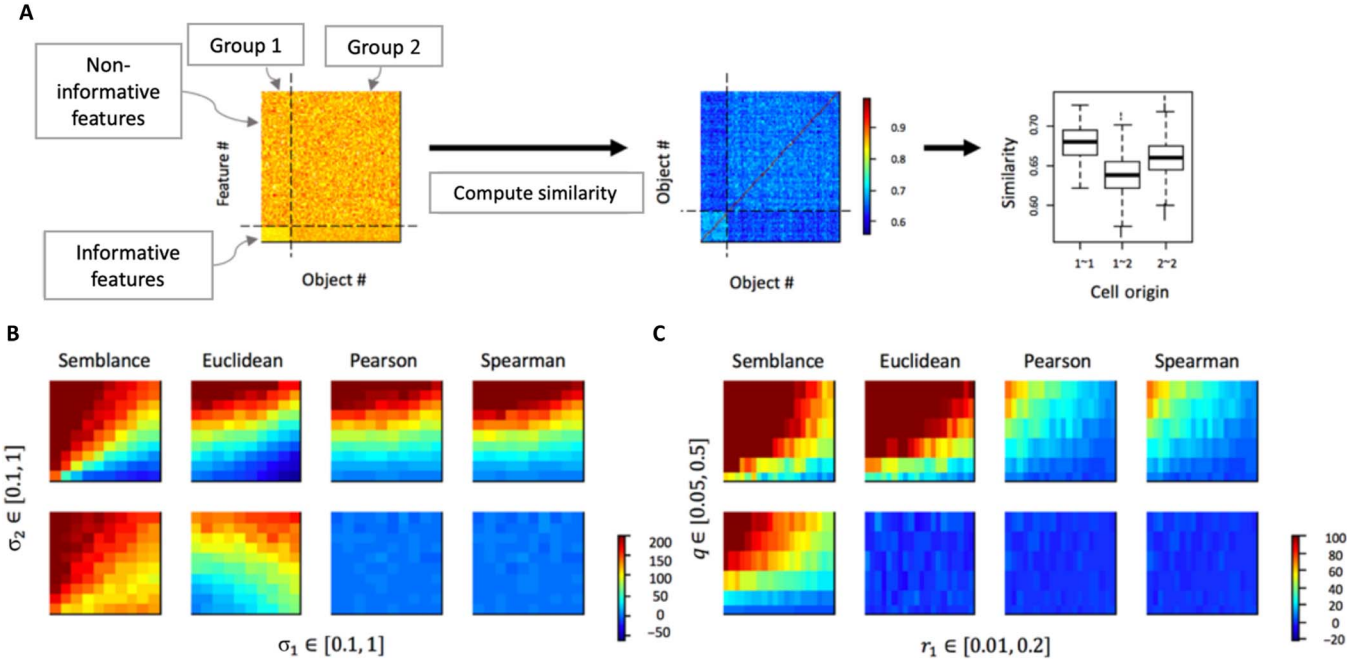


Fig. 2. Simulations exploring the effectiveness of similarity/distance measures. (**A**) Setup for one simulation run. (**B**) T_1 (top) and T_2 (bottom) values for each similarity/distance metric, for varying values of $\sigma_1 \in [0.1, 1]$ (horizontal axis) and $\sigma_2 \in [0.1, 1]$ (vertical axis). (**C**) T_1 (top) and T_2 (bottom) values for each similarity/distance metric, for varying values of $r_1 \in [0.1, 1]$ (horizontal axis) and $q \in [0.1, 1]$ (vertical axis).

following Eq. 7 with $\mu = 2$ and σ_1, σ_2 varying from 0.1 to 1. Heatmaps in the top row show the values of T_1, and those in the bottom row show the values of T_2 for each of the four similarity/distance measures. We see that Semblance improves upon Euclidean distance, Pearson, and Spearman, attaining large values for T_1 and T_2 across a broad range of parameters, especially when σ_2 is small. Figure 2C shows another set of simulations, with the same n, m, and p values as Fig. 2B, but under the model (Eq. 8) with $r_2 = 0.5$, r_1 varying from 0.01 to 0.2, and q varying from 0.05 to 0.5. We see that, in this case, there is no signal in T_2 for all of the measures except Semblance, and both Pearson correlation and Spearman correlation fail to separate the two groups for much of the parameter range. In contrast, Semblance gives large values for both T_1 and T_2 for a large portion of the explored parameter region.

We explored varying combinations of p, q, σ_1, and σ_2 in the normal setting, and p, q, r_0, and r_1 in the Bernoulli setting. Summarizing these systematic experiments in representative heatmaps (Fig. 3), we found that Semblance has robust performance across different distributions and distribution parameters (σ_1, σ_2, r_1, and r_2) as long as the proportion of informative features is not too small. Semblance is better than the other metrics especially in differentiating small tight subpopulations, i.e., niche groups. Unweighted Semblance ($w_g = 1$ in Eq. 2) retains less information and should not be used when informative features are extremely rare ($p \to 0$), but the separation between clusters is extremely large ($p \to 0$, $\mu \to \infty$). This lack of sensitivity for rare features, however, can be remedied by the use of distribution-informative weights w_g.

In its explicit construction, Semblance is shift and scale invariant; however, this robustness comes at a trade-off of being insensitive to the shape of the distribution. We compared the performance of a Semblance metric that does not put explicit weights on the features ($w_g = 1$) with a modified metric where features are weighed on the basis of the shape of the distribution ($w_g \neq 1$ in Eq. 2). We used as weights the negentropy, a robust approximation of kurtosis and non-Gaussianity that is invariant for invertible linear transformations of data (*13*), in our simulations described above and found that negentropy weighting further enhances the information captured by Semblance (fig. S1, A to C). For positive-valued data, we recommend the use of Gini coefficient for weights. The Gini index is a robust measure of dispersion (*14*). It ranges from 0 to 1, wherein a value of 0 means that values of the feature are distributed perfectly uniformly across the objects, and a value close to 1 means that there is high dispersion of values. When the data are simulated from gamma distributions (fig. S1D), we found that Gini weighting provides a robust and sensitive method of feature selection. Compared to an unweighted Semblance, the Gini-weighted statistic carries more signal, especially when the fraction of informative features is small (fig. S1, E to G). Collectively, our simulations demonstrate that Semblance can perform well in a totally unsupervised fashion; nonetheless, when prior knowledge about the data distribution is available, Semblance can also incorporate that to weigh features and augment its performance.

Semblance kernel-tSNE identifies a niche retinal horizontal cell population

In the setting of single-cell RNA sequencing (scRNA-seq), the data are in the form of a matrix with each row representing a cell and each column representing a gene. For cell c and gene g, N_{cg} is a count matrix

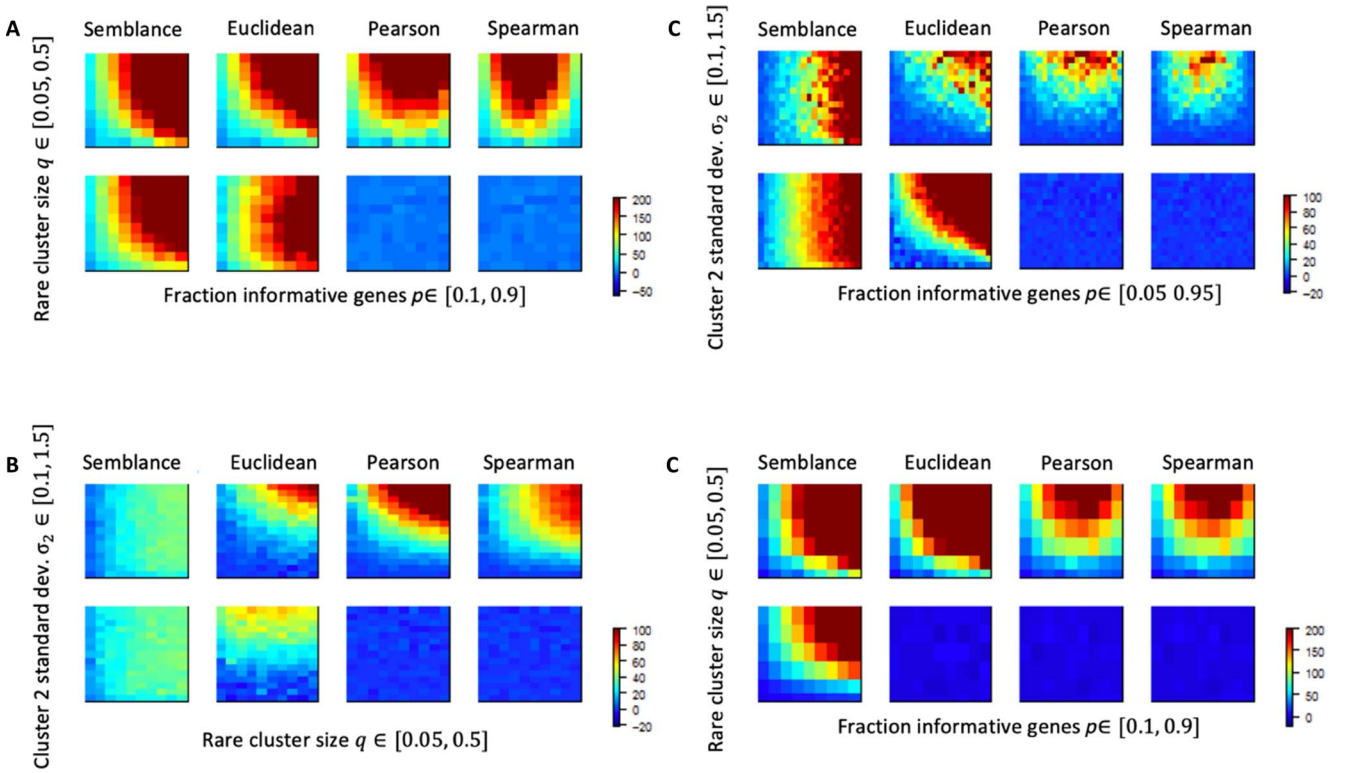
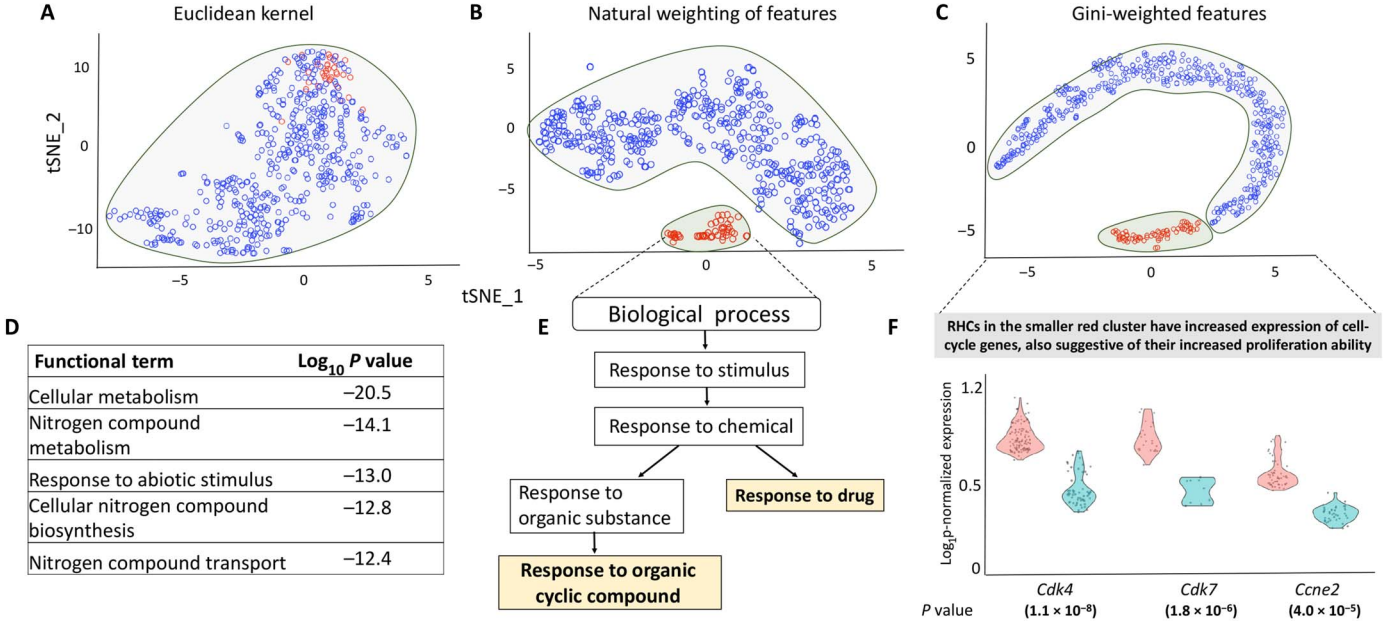


Fig. 3. Simulation results over parameter sweeps. For each 2 × 4 group of heatmaps, the top row shows T_1 and the bottom row shows T_2 for each similarity/distance metric, computed as described in the text. Simulation parameters are varied along the rows and columns of the heatmaps. (**A**) Normal model, $p = \{0.1, 0.2, \ldots, 0.9\}$ for the horizontal axis, and $q \in \{0.05, 0.1, \ldots, 0.5\}$ for the vertical axis. (**B**) Normal model, $q \in \{0.05, 0.1, \ldots, 0.5\}$ for the horizontal axis and $\sigma_2 \in \{0.1, 0.2, \ldots, 1.5\}$ for the vertical axis. (**C**) Normal model, $p = \{0.1, 0.2, \ldots, 0.9\}$ for the horizontal axis and $\sigma_2 \in \{0.1, 0.2, \ldots, 1.5\}$ for the vertical axis. (**D**) Binomial model, $p = \{0.1, 0.2, \ldots, 0.9\}$ for the horizontal axis, and $q \in \{0.05, 0.1, \ldots, 0.5\}$ for the vertical axis.

measuring a gene's RNA expression level in the given cell. A first step in the analysis of these data is often visualization via a t-distributed stochastic neighbor embedding (tSNE)–type dimension reduction. Most studies arbitrarily use the Euclidean distance or the radial basis function (RBF) in this step, although methods based on more sophisticated kernel choices that rely on strong prior assumptions have been proposed (*15*). Starting from the low-dimension embedding, a primary goal in many single-cell studies is to classify cells into distinct cell types and identify previously unknown cell subpopulations. This is a challenging analysis due to many factors: (i) Expression levels are not comparable across genes—lowly expressed cell-type markers may be swamped by highly expressed housekeeping genes; (ii) gene expression at the single-cell level is often bursty and thus cannot be approximated by the normal distribution; (iii) one is often interested in detecting rare niche subpopulations for which current methods have low power. These considerations motivated us to use the Semblance kernel to compute a cell-to-cell similarity metric, which can be used as input to tSNE, principal components analysis (PCA), and other kernel-based algorithms. Most methods used for cell-type identification based on scRNA-seq limit their consideration to highly variable genes, thereby using only a subset of the features. Instead, Semblance can be computed over all features, ensuring that information from all informative genes is retained.

Consider the retinal horizontal cell (RHC), a unique cell type that recently came to the limelight because of its notable morphological plasticity, and its role as a possible precursor for retinoblastoma (*16*). RHCs have a special level of complexity wherein they can undergo migration, mitosis, and differentiation at late developmental stages. They are traditionally divided into H1 axon-bearing and H2 to H4 axonless subtypes, although the latter are largely absent in the rod-dominated retina of most mammals (*17*). The axon-bearing and axonless RHC subtypes are generated during retinal development from progenitors that are susceptible to a transition in metabolic activity. For example, follistatin, an anabolic agent that alters protein synthesis and the inherent metabolic architecture in tissues, increases RHC proliferation (*18*). RHC subtypes also exhibit temporally distinguishable periods of migration, likely affected by their cellular metabolic state. These distinctive features are controlled by a niche set of genes, and thus RHCs provide a nonpareil setting to test Semblance. We used unweighted Semblance on an scRNA-seq dataset of 710 Lhx1^{+} RHCs from healthy P14 mice (*19*) and sought to answer the question, How similar are RHCs to each other? When we use Euclidean distance for tSNE analysis, only one RHC cluster could be identified (Fig. 4A), as opposed to two subsets of RHCs identified using kernel-tSNE (Fig. 4B).

Moreover, since the Gini coefficient is a robust measure of gene expression dispersion in single-cell transcriptomics experiments (*20*), we wondered how the unweighted Semblance-tSNE ($w_g = 1$ in Eq. 2) would compare against a Semblance measure where features are Gini-weighted. Although both kernel-tSNE projections successfully separated the second/rare RHC cluster, weighing genes by their Gini coefficient also suggested an underlying geometry, pointing to a plausible trajectory (Fig. 4C and fig. S2). We then sought further biological interpretation of these results and discovered that the cells in the second, smaller cluster—comprising 12% of the total RHC population—identified by Semblance exhibit differential expression (DE) of pathways that affect metabolism (Fig. 4D). We explicated our results by testing for enriched Gene Ontology (GO) functional categories using REVIGO (*21*) and uncovered a niche RHC population that has unique metabolic response properties (Fig. 4E). Gini weighting showed that this niche RHC cohort, which



Functional term	Log_{10} *P* value
Cellular metabolism	−20.5
Nitrogen compound metabolism	−14.1
Response to abiotic stimulus	−13.0
Cellular nitrogen compound biosynthesis	−12.8
Nitrogen compound transport	−12.4

Fig. 4. A unique RHC cluster is identified by Semblance kernel-tNSE. Each dot in (A) to (C) represents a single cell. Euclidean distance tSNE identifies a single RHC cluster (**A**) as opposed to two subpopulations identified by Semblance. Comparing the kernel's performance when features are naturally weighted on skewness (**B**) versus when they are weighted based on Gini coefficient (**C**) points to a geometric, trajectory-like structure in the data. The top five pathways found to be enriched in the rare cellular subtype are shown (**D**), and GO analysis suggested that the smaller RHC cluster has unique metabolic response properties (**E**). We also found evidence that these metabolic properties might lead to increased proliferation as suggested by increased expression of cell cycle genes by the cells in the red/rare cluster (**F**). For DE analysis, Benjamini-Hochberg–corrected *P* values are noted underneath each cyclin gene; the color codes blue and red correspond to the major and rare RHC clusters, respectively.

has an up-regulated metabolic state, resides at one end of a trajectory, suggesting that this trajectory may be related to proliferation. In other words, the niche RHC cohort could be a group of proliferating horizontal cells, compared with the more mature RHCs that constitute the larger cell cluster. To examine this, we tested for DE of cell cycle–associated genes, which are common markers of proliferation (*22*), between the two groups using MAST (model-based analysis of single-cell transcriptomics) (*23*). We found evidence for DE of genes associated with increased proliferation ability (Fig. 4F) in support of our hypothesis. Thus, without any domain knowledge, exploratory analysis using Semblance successfully uncovered meaningful biological signals from these data.

Semblance kernel PCA is efficient at image reconstruction and compression

Kernel PCA (kPCA), the nonlinear version of PCA, exploits the structure of high-dimensional features and can be used for data denoising, compression, and reconstruction (*24*, *25*). This task, however, is nontrivial because the kPCA output resides in some high-dimensional feature space and does not necessarily have preimages in the input space (*26*). kPCA, particularly using the Gaussian kernel defined by $k(x, y) = \exp(-||x - y||^2/2\sigma^2)$, has been used extensively to improve active shape models, reconstruct preimages, and recreate compressed shapes owing to its ability to recognize more nuanced features in real-world pictures (*27*). Nonlinear data recreation based on kPCA rests on the principle that using a small set of some *f* kPCA features provides an *f*-dimensional reparametrization of the data that better captures its inherent complexity (*28*). Since Semblance is nonparametric and empirically driven, emphasizing rare or niche feature values in data, we surmised that it would be useful as a nonlinear image reconstruction method. We discovered that Semblance kPCA can be used to reconstruct real-world images with remarkably good performance (Fig. 5, A and B). Upon adding uniform noise to an image, we found that Semblance kPCA can denoise images and compares favorably against linear PCA and Gaussian kPCA (Fig. 5, C and D).

We further evaluated the performance of kPCA on pictures obtained from The Yale Face Database (http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html) and the Bioconductor package EBImage (*29*) and found that Semblance can give a good re-encoding of the data when it lies along a nonlinear manifold, as is often the case with images. In each experiment, we computed the projections of the given image data onto the first *f* components and then sought to reconstruct the image as precisely as possible. We found that Semblance kPCA performed better than linear PCA and Gaussian kPCA when using a comparable number of components (fig. S3). This encouraging observation is supported by the intuitions underlying the construction of Semblance. Linear PCA encapsulates the coarse data structure as well as the noise. In contrast, Gaussian kPCA, similar to a *k*-nearest neighbor method, recreates the

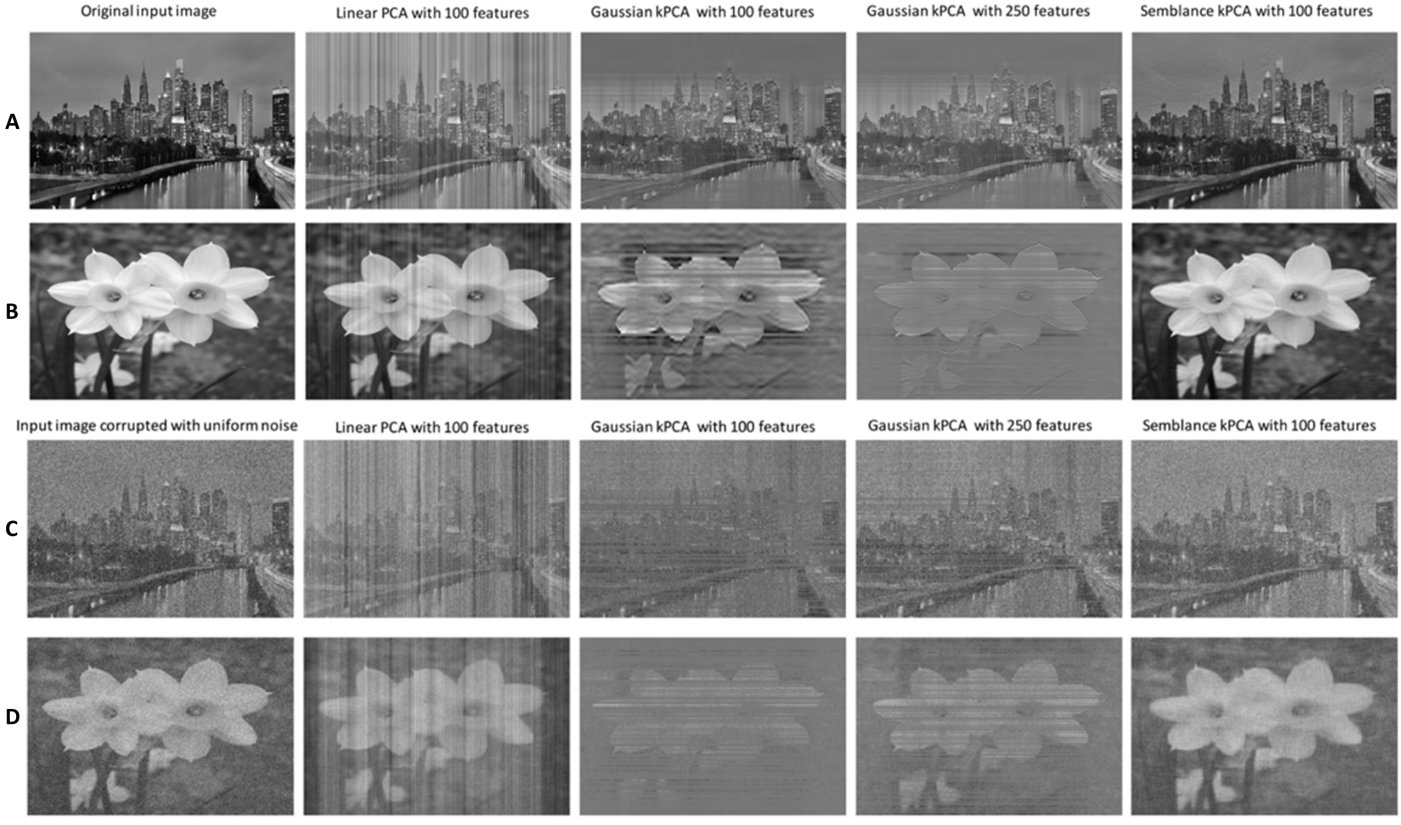


Fig. 5. kPCA using the Semblance kernel provides a useful method for image reconstruction and denoising. Two examples of open-source images: Philadelphia skyline (**A**) and daffodil flowers (**B**) are shown here. Semblance kPCA was able to effectively recover and compress when compared with linear PCA or Gaussian kPCA. These images were corrupted with added uniform noise: (**C**) and (**D**), respectively. The recovered image output using linear PCA, Gaussian kPCA, and Semblance kPCA is displayed. Comparing the same number of features (and even 2.5× as many features for Gaussian kPCA), Semblance performs favorably. More examples are given in the Supplementary Materials. Photo credits: Mo Huang (The Wharton School) and the EB Image Package.

connection between data points that are close to each other in the original feature space (*30*). On the other hand, Semblance apprehends data points that are proximal in the feature space under the metric determined by the empirical data distribution and therefore performs better when informative and non-informative features in an image are not on the same scale.

Semblance performs comparably with domain-specific kernel support vector machines in stock market forecasting

In finance and business analytics, although predictions of market volatility are inherently challenging, support vector machines (SVMs) using Gaussian and Laplacian kernels have been found to efficiently model stock market prices, partly because training these kernel SVMs (kSMVs) allows convex optimization with a linear constraint, resulting in a stable and unique global minimum (*31*, *32*). Semblance is a context-free kernel, and therefore one might be tempted to rely on existing kernels that explicitly incorporate domain-specific information. To examine Semblance in the setting of financial forecasting, we compared the performance of Semblance against eight other kernel functions. We used the Center for Research in Security Prices (CRSP) Database (http://www.crsp.com/products/research-products/crspziman-real-estate-database), which combines stock price and returns data with financial indices and company-specific information on all real estate investment trusts (REITs) that have traded on the three primary exchanges: Nasdaq, New York Stock Exchange (NYSE), and NYSE MKT. We focused our analysis on the 5419 REITs that were actively trading between 1 January 2016 and 31 December 2017 and obtained a list of financial indices and company-specific indicators for each REIT (table S1).

We determined which REITs had a net positive rate of return on their stock and then classified the companies based on whether they had a positive or negative rate of return. We then used kSVM to determine how accurately the model was able to predict the REIT category. We randomly split the data into training and test subsets in a 3:1 ratio and compared the generalization ability of each kSVM classifier using 10-fold cross-validation. Consistent with previous research in this area (*33*, *34*), we observed that the Gaussian and Laplacian kernels performed better than linear, spline, and hyperbolic tangent kernels likely because the former two kernels are homogeneous and have good approximation capabilities to model financial fluctuations. Nonetheless, Semblance was more accurate at REIT classification than most other kernel choices (table S2) and performed comparably to other popularly used context-specific kernels.

DISCUSSION

We have presented Semblance, a new similarity kernel for the analysis of multivariate data. Semblance relies on the simple intuition that the empirically observed distribution for each feature should be used to reward a premium to proximity among objects in low-probability regions of the feature space. In this way, Semblance is sensitive for detecting niche features in the data. We have shown that Semblance is a valid Mercer kernel and thus can be used in a principled way in kernel-based learning algorithms. From a computational point of view, Semblance enables the extraction of features of the data's empirical distribution at low computational cost. It naturally relies only on ranked feature values and thus is extremely robust. We evaluated Semblance and compared it to some commonly used similarity measures through simulations and diverse real-world examples, demonstrating scenarios where Semblance can improve downstream analysis.

Kernelized learning methods have been tremendously useful in a wide variety of applications and disciplines, particularly because of their ability to map data in complex, nonlinear spaces (*6*, *25*, *28*). Most commonly, kernels have been used to compare and classify objects, for instance, in clustering algorithms (*7*); however, Mercer kernels have another important interpretation in that they reflect similarities between sets of local features in the data. Semblance exploits this latter concept by defining a general-purpose similarity measure on probability spaces, even when no explicit correspondence between the data might appear obvious or intuitive. Satisfying the Mercer condition ensures that Semblance will guarantee unique global optimal solutions for downstream learning algorithms (*35*). Semblance operates in a high-dimensional, implicit feature space and can be applied to any data domain. We anticipate that it will also find utility in "multiple-kernel learning" approaches, wherein multiple kernels are often combined to learn from a heterogeneous data source.

MATERIALS AND METHODS

Algorithm to implement the semblance kernel

procedure STEP 1
For a given feature, g, create a descending ranked list such that the object with the highest value of g is ranked 1, and the object with lowest value is ranked last.
procedure STEP 2
Compute the empirical cumulative distribution function (CDF) for the feature g.
loop:
for features $g : 1 \rightarrow G$ **do**
Store lists as look-up tables
For a given feature g, determine where two observations, x and y, fall on the CDF for g.
procedure STEP 3
Calculate the difference between the ranks of x and y,
Add 1 to the difference between ranks.
procedure STEP 4
loop:
for features $g : 1 \rightarrow G$ **do**
Step 3, and store the cumulative sum in a matrix as entry (x, y), corresponding to the xth row and yth column

R package implementation

Semblance is an open-source R package available on CRAN (Comprehensive R Archive Network; https://cran.r-project.org/web/packages/Semblance/) and is compatible with existing kernel method libraries such as kernlab (*36*). In our R package, we implemented the kernel method in the *ranksem* function, which takes an input N_{ng} matrix (of g feature measurements for n objects), and returns an $n \times n$ similarity matrix.

Proofs concerning NND matrices

Lemma 1: The sum of NND matrices is NND.
Proof: Let $K_{g(1)}$ and $K_{g(2)}$ be two NND matrices, such that $\forall z \in \mathbb{R}^n$
$z^T K_{g(1)} z$ and $z^T K_{g(2)} z > 0 \Rightarrow z^T K_{g(1)} z + z^T K_{g(2)} z > 0$
Using the distributive law of matrix multiplication

$$0 < z^T K_{g(1)} z + z^T K_{g(2)} z = z^T (K_{g(1)} + K_{g(2)}) z$$
$$\Rightarrow z^T (K_{g(1)} + K_{g(2)}) z > 0 \therefore (K_{g(1)} + K_{g(2)}) \succ 0$$

Lemma 2: Permuting the observations of an NND matrix preserves the NND structure.

Proof: Let π be the permutation matrix such that it has exactly one entry in each row and in each column equal to 1, and all other entries are 0. For any permutation matrix, $\pi^{-1} = \pi^T$ and thus

$$\pi\pi^T = \pi^T\pi = I$$

For any given NND matrix, K, $\pi K\pi^T$ is also NND. $\pi K\pi^T$ is also symmetric as

$$w^T(\pi K\pi^T)w = (\pi^T w)^T K(\pi^T w) \quad \forall w \neq 0 \text{ since } K \text{ is NND}$$

Furthermore, every NND matrix can be factored as $K = A^TA$, where A is the Cholesky factor of K. The Cholesky factorization of NND matrices is numerically stable—a principal permutation of the rows and columns does not numerically destabilize the factorization (*37*). This leads to the result that symmetrically permuting the rows and columns of an NND matrix yields another NND matrix.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/12/eaau9630/DC1

Semblance and the connection to Mercer kernels

Existence of corresponding feature space for Semblance

Table S1. List of technical indicators recorded for each observation/REIT by the CRSP Real Estate Database.

Table S2. Test accuracy in forecasting whether the rate of return for an REIT would be positive or negative using SVMs for a range of kernel choices.

Fig. S1. Comparison of a naturally weighted Semblance metric with one wherein features are weighed by a context-dependent measure.

Fig. S2. We tested Semblance on an scRNA-seq dataset with 710 RHCs (*19*) and compared its performance against the conventionally used, Euclidean distance–based analysis.

Fig. S3. kPCA using the Semblance kernel is able to efficiently reconstruct images.

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. J. M. Gavilan, F. V. Morente, Three similarity measures between one-dimensional data sets. *Rev. Colomb. Estad.* **37**, 79 (2014).
2. B. Schweizer, A. Sklar, Statistical metric spaces. *Pacific J. Math.* **10**, 313–334 (1960).
3. S. K. M. Wong, Y. Y. Yao, paper presented at the Proceedings of the 10th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, LA, USA, 1987.
4. M. G. Genton, Classes of kernels for machine learning: A statistics perspective. *J. Mach. Learn. Res.* **2**, 299–312 (2002).
5. X. Wang, E. P. Xing, D. J. Schaid, Kernel methods for large-scale genomic data analysis. *Brief. Bioinform.* **16**, 183–192 (2015).
6. T. Hofmann, B. Schölkopf, A. J. Smola, Kernel methods in machine learning. *Ann. Stat.* **36**, 1171–1220 (2008).
7. J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*. (Cambridge Univ. Press, 2004), p. xiv, 462 pp.
8. A. Gisbrecht, A. Schulz, B. Hammer, Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing* **147**, 71–82 (2015).
9. S. Saitoh, *Theory of Reproducing Kernels and Its Applications*, Pitman research notes in mathematics series (Longman Scientific & Technical, Wiley, 1988), 157 pp.
10. D. J. Schaid, Genomic similarity and kernel methods I: Advancements by building on mathematical and statistical foundations. *Hum. Hered.* **70**, 109–131 (2010).
11. H. Q. Minh, P. Niyogi, Y. Yao, in *Mercer's Theorem, Feature Maps, and Smoothing* (Springer Berlin Heidelberg, 2006), pp. 154–168.
12. J. Liu, R. Huang, Generalized Schur complements of matrices and compound matrices. *Electron. J. Linear Al.* **21**, (2010).
13. K. J. Cios, R. W. Swiniarski, W. Pedrycz, L. A. Kurgan, in *Data Mining: A Knowledge Discovery Approach* (Springer US, 2007), pp. 133–233.
14. C. S. Moskowitz, V. E. Seshan, E. R. Riedel, C. B. Begg, Estimating the empirical Lorenz curve and Gini coefficient in the presence of error with nested data. *Stat. Med.* **27**, 3191–3208 (2008).
15. B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, S. Batzoglou, Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14**, 414–416 (2017).
16. R. A. Poché, B. E. Reese, Retinal horizontal cells: Challenging paradigms of neural development and cancer biology. *Development* **136**, 2141–2151 (2009).
17. H. Boije, S. Shirazi Fard, P. H. Edqvist, F. Hallbook, Horizontal cells, the odd ones out in the retina, give insights into development and disease. *Front. Neuroanat.* **10**, 77 (2016).
18. P.-H. Edqvist, M. Lek, H. Boije, S. M. Lindbäck, F. Hallböök, Axon-bearing and axon-less horizontal cell subtypes are generated consecutively during chick retinal development from progenitors that are sensitive to follistatin. *BMC Dev. Biol.* **8**, 46 (2008).
19. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, S. A. McCarroll, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
20. J. Wang, M. Huang, E. Torre, H. Dueck, S. Shaffer, J. Murray, A. Raj, M. Li, N. R. Zhang, Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E6437–E6446 (2018).
21. F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLOS ONE* **6**, e21800 (2011).
22. M. L. Whitfield, L. K. George, G. D. Grant, C. M. Perou, Common markers of proliferation. *Nat. Rev. Cancer* **6**, 99–106 (2006).
23. G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, R. Gottardo, MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278, 278 (2015).
24. A. K. Romney, C. C. Moore, W. H. Batchelder, T.-L. Hsia, Statistical methods for characterizing similarities and differences between semantic structures. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 518–523 (2000).
25. B. Scholkopf, A. J. Smola, K.-R. Muller, in *Advances in Kernel Methods*, S. Bernhard, J. C. B. Christopher, J. S. Alexander, Eds. (MIT Press, 1999), pp. 327–352.
26. S. Mika, B. Schölkopf, A. Smola, K. -B. Müller, M. Scholz, G. Rätsch, paper presented at the Proceedings of the 1998 conference on Advances in neural information processing systems II, 1999.
27. H. Lu, F. Yang, in *Subspace Methods for Pattern Recognition in Intelligent Environment*, Y.-W. Chen, L. C. Jain, Eds. (Springer Berlin Heidelberg, 2014), pp. 1–31.
28. S. Y. Kung, in *Kernel Methods and Machine Learning*, S. Y. Kung, Ed. (Cambridge Univ. Press, 2014), pp. 77–78.
29. G. Pau, F. Fuchs, O. Sklyar, M. Boutros, W. Huber, EBImage—An R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979–981 (2010).
30. K. Q. Weinberger, F. Sha, L. K. Saul, paper presented at the Proceedings of the twenty-first international conference on Machine learning, Banff, Alberta, Canada, 2004.
31. V. P. Upadhyay, S. Panwar, R. Merugu, R. Panchariya, paper presented at the Proceedings of the International Conference on Advances in Information Communication Technology & Computing, Bikaner, India, 2016.
32. L. J. Cao, F. H. Tay, Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans. Neural Netw.* **14**, 1506–1518 (2003).
33. A. Fan, M. Palaniswami, Selecting bankruptcy predictors using a support vector machine approach. *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, Como, Italy, 27 July 2000.
34. Z. Hu, J. Zhu, K. Tse, Stocks market prediction using Support Vector Machine, *6th International Conference on Information Management, Innovation Management and Industrial Engineering*, Xi'an, China, 23 to 24 November 2013.
35. B. Schölkopf, C. J. C. Burges, A. J. Smola, *Advances in Kernel Methods: Support Vector Learning* (MIT Press, 1999), p. vii, 376 pp.
36. A. Karatzoglou, A. Smola, K. Hornik, A. Zeileis, kernlab—An S4 Package for Kernel Methods in R 2004. *J. Stat. Softw.* **11**, 1–20 (2004).
37. S. Y. Stepanov, Symmetrization of the sign-definiteness criteria of symmetrical quadratic forms. *J. Appl. Math. Mech.* **66**, 933–941 (2002).

Acknowledgments: N.R.Z. thanks Z. Wu (Brown University, Rhode Island) for enlightening discussions. We thank H. Tang (Stanford University, California) for providing feedback on an earlier version of this research report. We also thank the referees for their helpful comments, which have led to a better presentation of the paper. **Funding:** This work was supported by the National Institutes of Health (NIH grant R01-GM125301 to N.R.Z.). **Author contributions:** D.A. and N.R.Z. devised the idea, conducted the supporting experiments, and wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests.

Data and materials availability: All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials and are publicly available on open-access repositories. Any additional computer codes related to this paper may be requested from the authors. The scRNA-seq data were generated using the Drop-Seq platform and are publicly accessible via the NCBI Gene Expression Omnibus (accession: GSE63473). The data on real estate investment trusts (REITs) are curated by the Center for Research in Security Prices (CRSP) and are accessible at www.crsp.com/products/research-products/crspziman-real-estate-database.

Submitted 14 September 2018
Accepted 30 September 2019
Published 4 December 2019
10.1126/sciadv.aau9630

Citation: D. Agarwal, N. R. Zhang, Semblance: An empirical similarity kernel on probability spaces. *Sci. Adv.* **5**, eaau9630 (2019).