Tom Collins – Computational Models of Translational Pattern Discovery

Supplemental Materials

Table S1 contains parameter values and corresponding short explanations for the first simulation, where point sets were generated using either Fourier descriptors [1] or scatter patterns (parameter G). In the case of Fourier descriptors, l points on a unit circle are displaced by sinusoids with randomly selected amplitudes, frequencies, and phases. The resulting points are scaled according to L_P (see Table S1) and the coordinates rounded to nearest integers to give one occurrence of a pattern. In the case of scatter patterns, points on a k-dimensional integer grid of side length L_P are randomly generated from a uniform distribution to give one occurrence of a pattern. For either a shape or scatter pattern, m translators are then generated on a k-dimensional integer grid of side length L_D using a Poisson random variable (see Table S1). In total, M patterns are generated per point set, where M follows a binomial distribution. Parameters in Table S1 were chosen such that the total number of pattern points was likely to remain below n, the desired size of the point set, whereupon the pattern points were interspersed with randomly selected points until reaching the desired set size. As n grew, Poisson and binomial parameters remained fixed, meaning the abstract shapes were interspersed with an increasing number of random points. This made it possible to study the effects of increasing set size and decreasing signal-to-noise ratio (SNR) on algorithm performance. For both shape and scatter conditions, sets were also generated that varied in size whilst keeping an approximately constant ratio of pattern points to point set size. Results reported in the main text focus on abstract shape data for a decreasing ratio of pattern points to point set size. The precision, recall, runtime, and FFTP trends in the other three conditions (shape, constant ratio; scatter, decreasing ratio; scatter, constant ratio) were largely identical to these results, demonstrating that the findings generalize beyond abstract shape data to other simulated point sets. Rows for r, a, b, and region in Table S1 (just below the line) refer to parameter settings for the algorithms SIAR and SIARCT.

For the generation of an image stack, images were imported in JPG format and converted from color to black and white (i.e., a matrix containing ones and zeros) using a thresholding technique such that the number of ones and zeros was close to equal. Points were defined by the nonzero matrix indices and one hundred were sampled at random without replacement to encode the image. A stack of fifty images was generated by sampling with replacement from a total of thirty encoded images, with the point set encodings being placed one behind the other to create a three-dimensional point set, referred to as an image stack.

For forecasting comovements in commodity price data, prices for each commodity were normalized independently to z-scores. They were then quantized by mapping to the nearest element in the sequence Ci/f_q, $i = -f_q, -f_q + 1\ldots, -1, 0, 1, 2,\ldots, f_q$, where C is the maximum absolute score and $f_q = 30$ is the quantization fidelity. It is worth mentioning that the number of translational patterns occurring in the quantized point set is a monotonic nondecreasing function of the quantization fidelity.

References

[1] C.T. Zahn and R.Z. Roskies, "Fourier Descriptors for Plane Closed Curves," *IEEE Trans. Computers*, vol. C-21, vol. 3, pp. 269-281, 1972.

TABLE S1
SIMULATION PARAMETERS (POINT SETS AND ALGORITHMS)

Parameter	Explanation	Current simulation value(s)	In general
k	Dimension of point set	$k = 2$	$k \geq 1$
n	Size of point set	$n \in \{1000, 1200, \ldots, 2000\}$	$n \geq 1$
R	Ratio of pattern points to set size (approximately decreasing or constant)	$R \in$ {"decreasing", "constant"}	Any
M	Number of patterns per point set, follows a binomial distribution with parameters governed by R	$M \sim \{B(25, 0.5), B(n/40, 0.5)\}$	Any
m	Number of occurrences of each pattern takes the maximum of a Poisson random variable and 2	$m \sim \max\{\text{Poiss}(5), 2\}$	Any
l	Point set size for each pattern follows a binomial distribution	$l \sim B(20, 0.5)$	Any
L_D	Side length of the k-dimensional integer grid on which points are generated	$L_D = 1000$	$L_D \geq 1$
L_P	Side length of the k-dimensional integer grid on which patterns are generated	$L_P = 30$	$1 \leq L_P \leq L_D$
G	Generation method (abstract shapes from Fourier descriptors or scatter patterns).	$G \in$ {"shape", "scatter"}	Any
S	Number of simulations per point set size	$S = 20$	$S \geq 1$
r	Number of superdiagonals used in SIAR and SIARCT	$r = 1$	$1 \leq r < n$
a	Compactness threshold used by the compactness trawler	$a = .95$	$a \in (0, 1]$
b	Points threshold used by the compactness trawler	$b = 5$	$b \geq 1$
region	Region type used to calculate compactness (lexicographic or convex hull)	region $\in$ {"LX", "CH"}	Any