Dimensionality reduction for classification of stochastic texture images

C.T.J. Dodson[1] and W.W. Sampson[2]
[1]School of Mathematics, [2]School of Materials,
University of Manchester, M13 9PL, UK
ctdodson@manchester.ac.uk and *william.sampson@manchester.ac.uk*

Abstract

Stochastic textures yield images representing density variations of differing degrees of spatial disorder, ranging from mixtures of Poisson point processes to macrostructures of distributed finite objects. They arise in areas such as signal processing, molecular biology, cosmology, agricultural spatial distributions, oceanography, meteorology, tomography, radiography and medicine. The new contribution here is to couple information geometry with multidimensional scaling, also called dimensionality reduction, to identify small numbers of prominent features concerning density fluctuation and clustering in stochastic texture images, for classification of groupings in large datasets. Familiar examples of materials with such textures in one dimension are cotton yarns, audio noise and genomes, and in two dimensions paper and nonwoven fibre networks for which radiographic images are used to assess local variability and intensity of fibre clustering. Information geometry of trivariate Gaussian spatial distributions of mean pixel density with the mean densities of its first and second neighbours illustrate features related to sizes and density of clusters in stochastic texture images. We derive also analytic results for the case of stochastic textures arising from Poisson processes of line segments on a line and rectangles in a plane. Comparing human and yeast genomes, we use 12-variate spatial covariances to capture possible differences relating to secondary structure. For each of our types of stochastic textures: analytic, simulated, and experimental, we obtain dimensionality reduction and hence 3D embeddings of sets of samples to illustrate the various features that are revealed, such as mean density, size and shape of distributed objects, and clustering effects.

Keywords: Dimensionality reduction, stochastic texture, density array, clustering, spatial covariance, trivariate Gaussian, radiographic images, genome, simulations, Poisson process

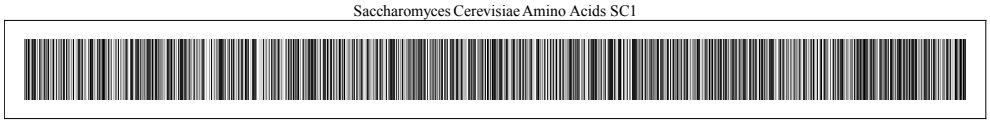

Figure 1: *Example of a 1-dimensional stochastic texture, a grey level barcode for the amino acid sequence in a sample of the Saccharomyces Cerevisiae yeast genome from the database [19].*

1 Introduction

The new contribution in this paper is to couple information geometry with dimensionality reduction, to identify small numbers of prominent features concerning density fluctuation and clustering in stochastic texture images, for classification of groupings in large datasets. Our methodology applies to any stochastic texture images, in one, two or three dimensions, but to gain an impression of the nature of examples we analyse some familiar materials for which we have areal density arrays, and derive analytic expressions of spatial covariance matrices for Poisson processes of finite objects in one and two dimensions. Information geometry provides a natural distance structure on the textures via their spatial covariances, which allows us to obtain multidimensional scaling or dimensionality reduction and hence 3D embeddings of sets of samples. See Mardia et al. [14] for an account of the original work on multidimensional scaling.

The simplest one-dimensional stochastic texture arises as the density variation along a cotton yarn, consisting of a near-Poisson process of finite length cotton fibres on a line, another is an audio noise drone consisting of a Poisson process of superposed finite length notes or chords. A fundamental microscopic 1-dimensional stochastic process is the distribution of the 20 amino acids along protein chains in a genome [1, 3]. Figure 1 shows a sample of such a sequence of the 20 amino acids A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y mapped onto the 20 grey level values $0.025, 0.075, \ldots, 0.975$ from the database [19], so yielding a grey-level barcode as a 1-dimensional texture. We analyse such textures in §6.5.

The largest 3-dimensional stochastic structure is the cosmological void distribution, which is observable via radio astronomy [1]. More familiar three-dimensional stochastic porous materials include metallic (Figure 2) and plastic solid foams, geological strata and dispersions in gels, observable via computer tomography [1]. Near-planar, non-woven stochastic

Figure 2: *Aluminium foam with a narrow Gaussian-like distribution of void sizes of around 1cm diameter partially wrapped in fragmented metallic shells, used as crushable buffers inside vehicle bodies. The cosmological void distribution is by contrast gamma-like with a long tail [8], interspersed with 60% of galaxies in large-scale sheets, 20% in rich filaments and 20% in sparse filaments [12]. Such 3D stochastic porous materials can both be studied by tomographic methods, albeit at different scales by different technologies, yielding sequences of 2D stochastic texture images.*

fibre networks are manufactured for a variety of applications such as, at the macroscale for printing, textiles, reinforcing, and filtration and at the nanoscale in medicine. Figure 3 shows a selection of electron micrographs for networks at different scales. Radiography or optical densitometry yield areal density images of the kinds shown in Figure 4.

Much analytic work has been done on modelling of the statistical geometry of stochastic fibrous networks [7, 1, 6, 17]. Using complete sampling by square cells, their areal density distribution is typically well represented by a log-gamma or a (truncated) Gaussian distribution of variance that decreases monotonically with increasing cell size; the rate of decay is dependent on fibre and fibre cluster dimensions. They have gamma void size distributions with a long tail. Clustering of fibres is well-approximated by Poisson processes of Poisson clusters of differing density and size. An unclustered Poisson process of single fibres is the standard reference structure for any given size distribution of fibres; its statistical geometry is well-understood for finite and infinite fibres. Note that any skewness associated with the underlying point process of fibre centres becomes negligible through the process of sampling by square cells [18].

Many stochastic textures arise from spatial processes that may be approximated by mixtures of Poisson or other distributions of finite objects or clusters of objects, in an analogous way to that which has been used for the past fifty years for the study of fibre networks. The Central

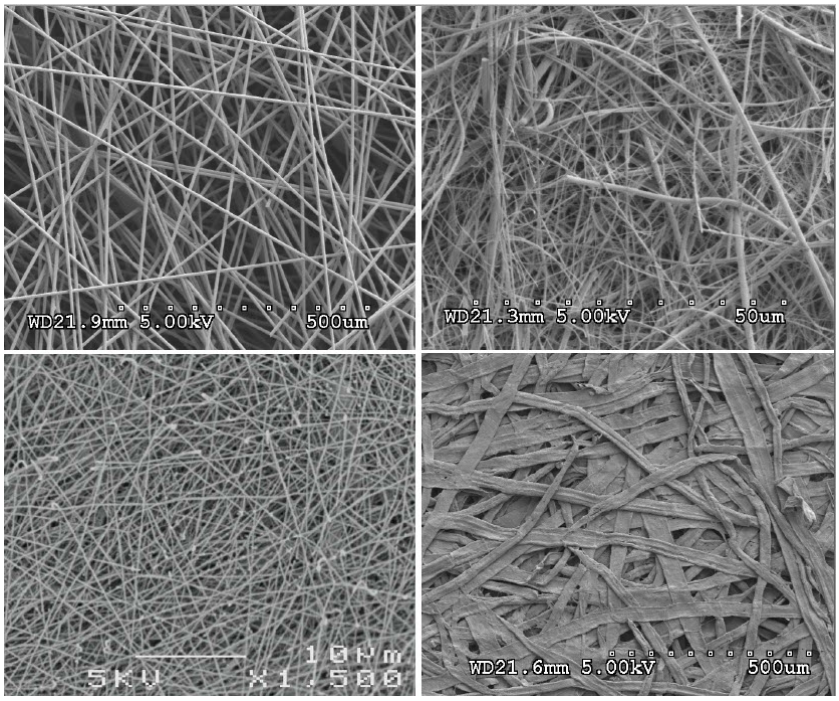

Figure 3: *Electron micrographs of four stochastic fibrous materials. Top left: Nonwoven carbon fibre mat; Top right: glass fibre filter; Bottom left: electrospun nylon nanofibrous network (Courtesy S.J. Eichhorn and D.J. Scurr); Bottom right: paper using wood cellulose fibres—typically flat ribbonlike, of length 1 to 2mm and width 0.02 to 0.03mm.*

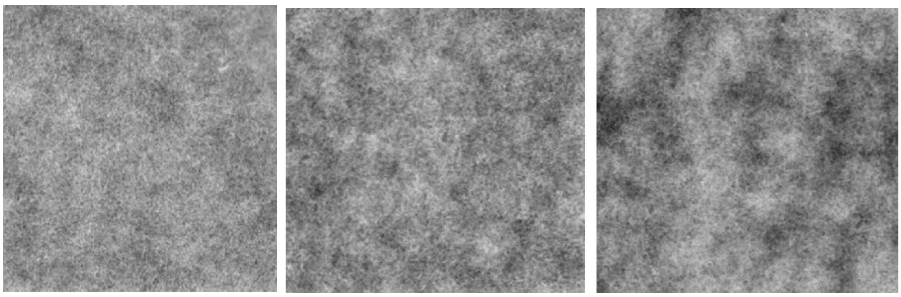

Figure 4: *Areal density radiographs of three paper networks made from natural wood cellulose fibres, with constant mean coverage,* $\bar{c} \approx 20$ *fibres, but different distributions of fibres. Each image represents a square region of side length 5 cm; darker regions correspond to higher coverage. The left image is similar to that expected for a Poisson process of the same fibres, so typical real samples exhibit clustering of fibres.*

Limit Theorem suggests that often such spatial processes may be represented by Gaussian pixel density distributions, with variance decreasing as pixel size increases, the gradient of this decrease reflecting the size distributions and abundances of the distributed objects and clusters, hence indicating the appropriate pixel size to choose for feature extraction. Once a pixel size has been chosen then we are interested in the statistics of the the three random variables: the mean density in such pixels, and the mean densities of its first and second neighbouring pixels. The correlations among these three random variables reflect the size and distribution of density clusters; this may be extended to more random variables by using also third, fourth, etc neighbours. In some cases, of course, other pixel density distributions may be more appropriate, such as mixtures of Gaussians.

2 Spatial covariance

The mean of a random value p is its average value, $\bar{p}$, over the population. The covariance $\mathrm{Cov}(p,q)$ of a pair of random variables, p and q is a measure of the degree of association between them, the difference between their mean product and the product of their means:

$$\mathrm{Cov}(p,q) = \overline{p\,q} - \bar{p}\,\bar{q}\ . \tag{1}$$

In particular, the covariance of a variable with itself is its variance. From the array of local average pixel density values $\tilde{\beta}_i$, we generate two numbers associated with each: the average density of the 6 first-neighbour pixels, $\tilde{\beta}_{1,i}$ and the average density of the 16 second-neighbour pixels, $\tilde{\beta}_{2,i}$. Thus, we have a trivariate distribution of the random variables $(\tilde{\beta}_i, \tilde{\beta}_{1,i}, \tilde{\beta}_{2,i})$ with $\bar{\beta}_2 = \bar{\beta}_1 = \bar{\beta}$.

Figure 5 provides an example of a typical data set obtained from a radiograph of a 5cm square commercial newsprint sample; the histogram and three-dimensional scatter plot show data obtained for pixels of side 1mm.

From the Central Limit Theorem, we expect the marginal distributions of $\tilde{\beta}_i$, $\tilde{\beta}_{1,i}$ and $\tilde{\beta}_{2,i}$ to be well approximated by Gaussian distributions. For the example in Figure 5, these Gaussians are represented by the solid lines on the histogram; this Gaussian approximation holds for all samples investigated in this study.

We have a simulator for creating stochastic fibre networks [10]. The code works by dropping clusters of fibres within a circular region where the centre of each cluster is distributed as a point Poisson process in the plane and the number of fibres per cluster, n_c, is a Poisson distributed

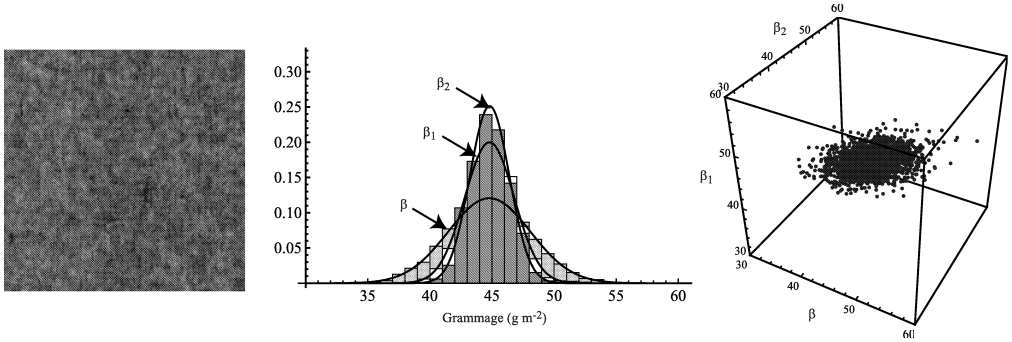

Figure 5: *Trivariate distribution of pixel density values for radiograph of a 5cm square newsprint sample. Left: source density map; centre: histogram of* $\tilde{\beta}_i$, $\tilde{\beta}_{1,i}$ *and* $\tilde{\beta}_{2,i}$*; right: 3D scatter plot of* $\tilde{\beta}_i$, $\tilde{\beta}_{1,i}$ *and* $\tilde{\beta}_{2,i}$.

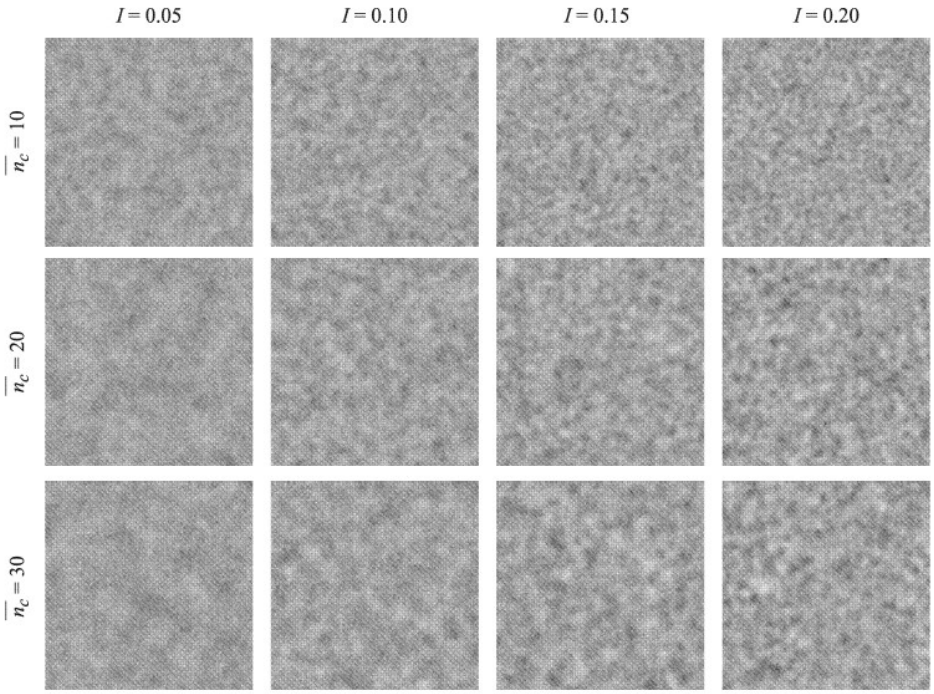

Figure 6: *Simulated areal density maps each representing a 4cm* $\times$ *4cm region formed from fibres with length* $\lambda = 1$ *mm, to a mean coverage of 6 fibres.*

random variable. The size of each cluster is determined by an intensity parameter, $0 < I \leq 1$ such that the mean mass per unit area of the cluster is constant and less than the areal density of a fibre. Denoting the length and width of a fibre by λ and ω respectively, the radius of a cluster containing n_c fibre centres is

$$r = \sqrt{\frac{n_c\,\lambda\,\omega}{\pi\,I}}\ . \tag{2}$$

Figure 6 shows examples of density maps generated by the simulator. We observe textures that increase in 'cloudyness' with n_c and increase in 'graininess' with I.

3 Analytic covariance for spatial Poisson processes of finite objects

Consider a Poisson process in the plane for finite rectangles of length λ and width $\omega \leq \lambda$, with uniform orientation of rectangle axes to a fixed direction. The covariance or autocorrelation function for such objects is known and given by [7]:
For $0 < r \leq \omega$

$$\alpha_1(r) = 1 - \frac{2}{\pi}\left(\frac{r}{\lambda} + \frac{r}{\omega} - \frac{r^2}{2\omega\lambda}\right). \tag{3}$$

For $\omega < r \leq \lambda$

$$\alpha_2(r) = \frac{2}{\pi}\left(\arcsin\left(\frac{\omega}{r}\right) - \frac{\omega}{2\lambda} - \frac{r}{\omega} + \sqrt{\frac{r^2}{\omega^2} - 1}\right). \tag{4}$$

For $\lambda < r \leq \sqrt{(\lambda^2 + \omega^2)}$

$$\begin{aligned}\alpha_3(r) \;=\; & \frac{2}{\pi}\left(\arcsin\left(\frac{\omega}{r}\right) - \arccos\left(\frac{\lambda}{r}\right) - \frac{\omega}{2\lambda} - \frac{\lambda}{2\omega} - \frac{r^2}{2\lambda\omega} + \right. \\ & \left. \sqrt{\frac{r^2}{\lambda^2} - 1} + \sqrt{\frac{r^2}{\omega^2} - 1}\right).\end{aligned} \tag{5}$$

Then, the coverage c at a point is the number of rectangles overlapping that point, a Poisson variable with grand mean value $\bar{c}$, and the average coverage or density in finite pixels $\tilde{c}$ tends to a Gaussian random variable. For sampling of the process using, say square inspection pixels of side length x, the variance of their density $\tilde{c}(x)$ is

$$Var(\tilde{c}(x)) = Var(c(0)) \int_0^{\sqrt{2}x} \alpha(r,\omega,\lambda)\ b(r)\ dr \tag{6}$$

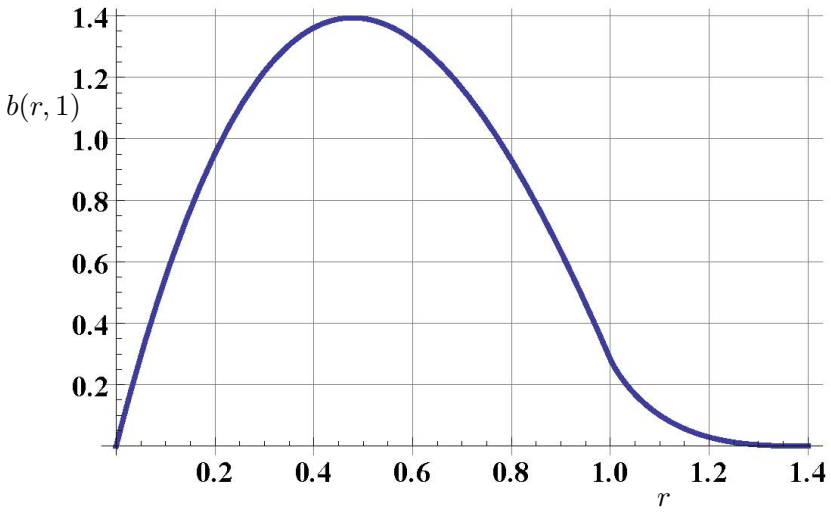

Figure 7: *Probability density function* $b(r, 1)$ *from equations (7),(8) for the distance* r *between two points chosen independently and at random in a unit square.*

where b is the probability density function for the distance r between two points chosen independently and at random in the given type of pixel; it was derived by Ghosh [13].
Using square pixels of side length x, for $0 \leq r \leq x$

$$b(r, x) = \frac{4r}{x^4}\left(\frac{\pi x^2}{2} - 2rx + \frac{r^2}{2}\right). \tag{7}$$

For $x \leq r \leq \sqrt{2}x$

$$\begin{aligned} b(r, x) \quad = \quad & \frac{4r}{x^4}\left(x^2\left(\arcsin\left(\frac{x}{r}\right) - \arccos\left(\frac{x}{r}\right)\right)\right) + \\ & \frac{4r}{x^4}\left(2x\sqrt{(r^2 - x^2)} - \frac{1}{2}(r^2 + 2x^2)\right). \end{aligned} \tag{8}$$

A plot of this function is given in Figure 7. Observe that, for vanishingly small pixels, that is points, b degenerates into a delta function on $r = 0$. Ghosh [13] gave also the form of b for other types of pixels; for arbitrary rectangular pixels those expressions can be found in [7]. For small values of r, so $r \ll D$, the formulae for convex pixels of area A and perimeter P all reduce to

$$b(r, A, P) = \frac{2\pi r}{A} - \frac{2Pr^2}{A^2}$$

which would be appropriate to use when the rectangle dimensions ω, λ are small compared with the dimensions of the pixel.

It helps to visualize practical variance computations by considering the case of sampling using large square pixels of side mx say, which themselves consist of exactly m^2 small square pixels of side x. The variance $Var(\tilde{c}(mx))$ is related to $Var(\tilde{c}(x))$ through the covariance $Cov(x, mx)$ of x-pixels in mx-pixels [7]:

$$Var(\tilde{c}(mx)) = \frac{1}{m^2} Var(\tilde{c}(x)) + \frac{m^2 - 1}{m^2} Cov(x, mx).$$

As $m \to \infty$, the small pixels tend towards points, $\frac{1}{m^2} Var(\tilde{c}(x)) \to 0$ so $Var(\tilde{c}(mx))$ admits interpretation as $Cov(0, mx)$, the covariance among points inside mx-pixels, the intra-pixel covariance, precisely $Var(\tilde{c}(mx))$ from equation (6).

The fractional between pixel variance for x-pixels is

$$\tilde{\rho}(x) = \frac{Cov(0, x)}{Var(c(0))} = \frac{Var(\tilde{c}(x))}{Var(c(0))}$$

which increases monotonically with λ and with ω but decreases monotonically with mx, see Deng and Dodson [6] for more details. In fact, for a Poisson process of rectangles the variance of coverage at points is precisely the mean coverage, $Var(c(0)) = \bar{c}$, so if we agree to measure coverage as a fraction of the mean coverage then equation (6) reduces to the integral

$$\frac{Var(\tilde{c}(x))}{\bar{c}} = \int_0^{\sqrt{2}x} \alpha(r, \omega, \lambda)\ b(r)\ dr = \tilde{\rho}(x). \tag{9}$$

Now, the covariance among points inside mx-pixels, $Cov(0, mx)$, is the expectation of the covariance between pairs of points separated by distance r, taken over the possible values for r in an mx-pixel; that amounts to the integral in equation (6). By this means we have continuous families of 2×2 covariance matrices for $x \in \mathbb{R}^+$ and $2 < m \in \mathbb{Z}^+$ given by

$$\begin{aligned} \Sigma^{x,m} &= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} Var(\tilde{c}(x)) & Cov(x, mx) \\ Cov(x, mx) & Var(\tilde{c}(x)) \end{pmatrix} \\ &= \begin{pmatrix} \tilde{\rho}(x) & \tilde{\rho}(mx) \\ \tilde{\rho}(mx) & \tilde{\rho}(x) \end{pmatrix}. \end{aligned} \tag{10}$$

which encodes information about the spatial structure formed from the Poisson process of rectangles, for each choice of rectangle dimensions $\omega \leq \lambda \in \mathbb{R}^+$. This can be extended to include mixtures of different rectangles with given relative abundances and processes of more complex objects such as Poisson clusters of rectangles.

There is a one dimensional version of the above that is discussed in [7, 6], with point autocorrelation calculated easily as

$$\alpha(r) = \begin{cases} 1 - \frac{r}{\lambda} & 0 \le r \le \lambda \\ 0 & \lambda < r. \end{cases} \tag{11}$$

Also, the probability density function for points chosen independently and at random with separation r in a pixel, which is here an interval of length x, is

$$b(r) = \frac{2}{x}(1 - \frac{r}{x}) \qquad (0 \le r \le x). \tag{12}$$

Then the integral (6) gives the fractional between pixel variance as

$$\tilde{\rho}(x, \lambda) = \begin{cases} 1 - \frac{x}{3\lambda} & 0 \le x \le \lambda \\ \frac{\lambda}{x}(1 - \frac{\lambda}{3x}) & \lambda < x. \end{cases} \tag{13}$$

So in the case of a one dimensional stochastic texture from a Poisson process of segments of length λ we have the explicit expression for the covariance matrices in equation (10):

$$\Sigma^{x,m}(\lambda) = \begin{pmatrix} \tilde{\rho}(x, \lambda) & \tilde{\rho}(mx, \lambda) \\ \tilde{\rho}(mx, \lambda) & \tilde{\rho}(x, \lambda) \end{pmatrix}. \tag{14}$$

In particular, if we take unit length intervals as the base pixels, for the Poisson process of unit length line segments, $x = \lambda = 1$ we obtain

$$\Sigma^{1,m}(1) = \begin{pmatrix} (1 - \frac{1}{3}) & \frac{1}{m}(1 - \frac{1}{3m}) \\ \frac{1}{m}(1 - \frac{1}{3m}) & (1 - \frac{1}{3}) \end{pmatrix} \quad \text{for } m = 2, 3, \ldots. \tag{15}$$

4 Information distance

Given the family of pixel density distributions, with associated spatial covariance structure among neighbours, we can use the Fisher metric [1] to yield an arc length function on the curved space of parameters which represent mean and covariance matrices. Then the information distance between any two such distributions is given by the length of the shortest curve between them, a geodesic, in this space. The computational difficulty is in finding the length of this shortest curve since it is the infimum over all curves between the given two points. Fortunately, in the cases we need, multivariate Gaussians, this problem has been largely solved analytically by Atkinson and Mitchell [2].

Accordingly, some of our illustrative examples use information geometry of trivariate Gaussian spatial distributions of pixel density with covariances among first and second neighbours to reveal features related to

sizes and density of clusters, which could arise in one, two or three dimensions. For isotropic spatial processes, which we consider here, the variables are means over shells of first and second neighbours, respectively. For anisotropic networks the neighbour sets would be split into more new variables to pick up the spatial anisotropy in the available spatial directions.

Other illustrations will use the analytic bivariate covariances given in the previous section §3 by equation (10).

What we know analytically is the geodesic distance between two multivariate Gaussians, A, B, of the same number n of variables in two particular cases [2]:

1. $\mu^{\mathbf{A}} \neq \mu^{\mathbf{B}}, \mathbf{\Sigma^A} = \mathbf{\Sigma^B} = \mathbf{\Sigma}$: $\quad f^A = (n, \mu^A, \Sigma), f^B = (n, \mu^B, \Sigma)$

$$D_\mu(f^A, f^B) = \sqrt{(\mu^A - \mu^B)^T \cdot \Sigma^{-1} \cdot (\mu^A - \mu^B)}. \tag{16}$$

2. $\mu^{\mathbf{A}} = \mu^{\mathbf{B}} = \mu, \mathbf{\Sigma^A} \neq \mathbf{\Sigma^B}$: $\quad f^A = (n, \mu, \Sigma^A), f^B = (n, \mu, \Sigma^B)$

$$D_\Sigma(f^A, f^B) = \sqrt{\frac{1}{2}\sum_{j=1}^{n} \log^2(\lambda_j)}, \tag{17}$$

$$\text{with } \{\lambda_j\} = \text{Eig}(\Sigma^{A^{-1/2}} \cdot \Sigma^B \cdot \Sigma^{A^{-1/2}}).$$

In the present paper we use equations (16) and (17) and take the simplest choice of a linear combination of both when mean and covariance are both different.

However, from the form of $D_\Sigma(f^A, f^B)$ in (17) we deduce that an approximate monotonic relationship arises with a more easily computed symmetrized log-trace function given by
$\Delta_\Sigma(f^A, f^B) =$

$$\sqrt{\log\left(\frac{1}{2n}\left(Tr(\Sigma^{A^{-1/2}} \cdot \Sigma^B \cdot \Sigma^{A^{-1/2}}) + Tr(\Sigma^{B^{-1/2}} \cdot \Sigma^A \cdot \Sigma^{B^{-1/2}})\right)\right)}. \tag{18}$$

This is illustrated by the plot in Figure 8 of $D_\Sigma(f^A, f^B)$ from equation (17) on $\Delta_\Sigma(f^A, f^B)$ from equation (18) for 185 trivariate Gaussian covariance matrices, where we see that

$$D_\Sigma(f^A, f^B) \approx 1.7\Delta_\Sigma(f^A, f^B).$$

A commonly used approximation for information distance is obtained from the Kullback-Leibler divergence, or relative entropy. Between two

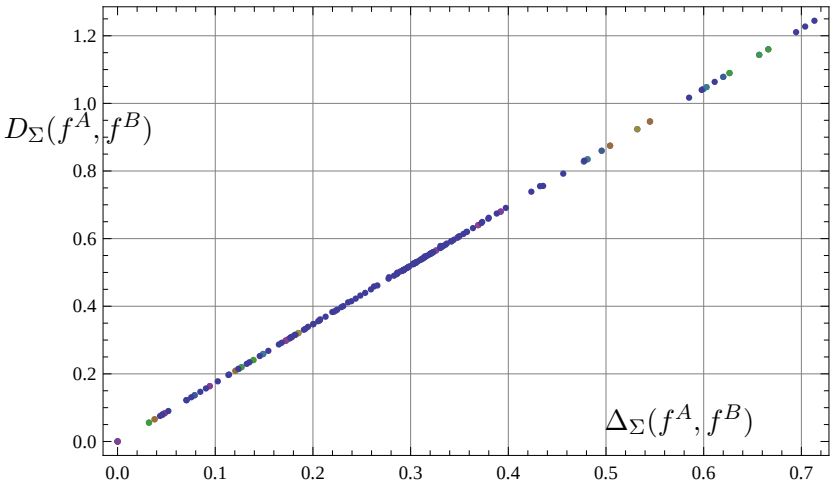

Figure 8: *Plot of* $D_\Sigma(f^A, f^B)$ *from (17) against* $\Delta_\Sigma(f^A, f^B)$ *from (18) for 185 different trivariate Gaussian covariance matrices.*

multivariate Gaussians $f^A = (n, \mu^A, \Sigma^A), f^B = (n, \mu^B, \Sigma^B)$ with the same number n of variables, its square root gives a separation measurement [16]:

$$\begin{aligned} KL(f^A, f^B) &= \frac{1}{2}\log(\frac{\det \Sigma^B}{\det \Sigma^A}) + \frac{1}{2}\mathrm{Tr}[\Sigma^{B^{-1}} \cdot \Sigma^A] \\ &+ \frac{1}{2}\left(\mu^A - \mu^B\right)^T \cdot \Sigma^{B^{-1}} \cdot \left(\mu^A - \mu^B\right) - \frac{n}{2}. \end{aligned} \tag{19}$$

This is not symmetric, so to obtain a distance we take the average KL-distance in both directions:

$$D_{KL}(f^A, f^B) = \sqrt{\frac{|KL(f^A, f^B)| + |KL(f^B, f^A)|}{2}} \tag{20}$$

The Kullback-Leibler distance tends to the information distance as two distributions become closer together; conversely it becomes less accurate as they move apart.

For comparing relative proximity, $\Delta_\Sigma(f^A, f^B)$ is a better measure near zero than the symmetrized Kullback-Leibler $D_{KL}(f^A, f^B)$ distance in those multivariate Gaussian cases so far tested and may be computationally quicker for handling large batch processes.

5 Dimensionality reduction of spatial density arrays

We shall illustrate the differences of spatial features in given data sets obtained from the distribution of local density for real and simulated planar stochastic fibre networks. In such cases there is benefit in mutual information difference comparisons of samples in the set but the difficulty is often the large number of samples in a set of interest—perhaps a hundred or more. Human brains can do this very well; the enormous numbers of optical sensors that stream information from the eyes into the brain with the result that we have a 3-dimensional reduction which serves to help us 'see' the external environment. We want to see a large data set organised in such a way that natural groupings are revealed and quantitative dispositions among groups are preserved. The problem is how to present the information contained in the whole data set, each sample yielding a 3×3 covariance matrix Σ and mean μ. The optimum presentation is to use a 3-dimensional plot, but the question is what to put on the axes.

To solve this problem we use multi-dimensional scaling, or dimensionality reduction, to extract the three most significant features from the set of samples so that all samples can be displayed graphically in a 3-dimensional plot. The aim is to reveal groupings of data points that correspond to the prominent characteristics; in our context we have different former types, grades and differing scales and intensities of fibre clustering. Such a methodology has particular value in the quality control for processes with applications that frequently have to study large data sets of samples from a trial or through a change in conditions of manufacture or constituents. Moreover, it can reveal anomalous behaviour of a process or unusual deviation in a product. The raw data of one sample from a study of spatial variability might typically consist of a spatial array of 250×250 pixel density values, so what we solve is a problem in classification for stochastic image textures.

The method, which we introduced in a preliminary report [11], depends on extracting the three largest eigenvalues and their eigenvectors from a matrix of mutual information distances among distributions representing the samples in the data set. The number in the data set is unimportant, except for the computation time in finding eigenvalues. This follows the methods described by Carter et al. [4, 5]. Our study is for datasets of pixel density arrays from complete sampling of density maps of stochastic textures which incorporate spatial covariances. We report the results of such work on a large collection of radiographs from commercial papers made from continuous filtration of cellulose and other fibres, [9].

The series of computational stages is as follows:

1. Obtain mutual 'information distances' $D(i,j)$ among the members of the data set of N textures $X_1, X_2, .., X_N$ using the fitted trivariate Gaussian pixel density distributions.

2. The array of $N \times N$ differences $D(i,j)$ is a real symmetric matrix with zero diagonal. This is centralized by subtracting row and column means and then adding back the grand mean to give $CD(i,j)$.

3. The centralized matrix $CD(i,j)$ is again a real symmetric matrix with zero diagonal. We compute its N eigenvalues $ECD(i)$, which are necessarily real, and the N corresponding N-dimensional eigenvectors $VCD(i)$.

4. Make a 3×3 diagonal matrix A of the first three eigenvalues of largest absolute magnitude and a $3 \times N$ matrix B of the corresponding eigenvectors. The matrix product $A \cdot B$ yields a $3 \times N$ matrix and its transpose is an $N \times 3$ matrix T, which gives us N coordinate values (x_i, y_i, z_i) to embed the N samples in 3-space.

Example: Bivariate Gaussians

$$
\begin{aligned}
f(x,y) &= \frac{1}{2\pi\sqrt{\Delta}} exp \frac{-1}{\Delta^2}(y-\mu_2)^2\sigma_{11} + (x-\mu_1)[(x-\mu_1)\sigma_{22} + 2(-y+\mu_2))\sigma_{12}] \\
\mu &= (\mu_1, \mu_2), \\
\Delta &= Det[\Sigma] = \sigma_{11}\sigma_{22} - \sigma_{12}^2 \\
\Sigma &= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = \sigma_{11}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \sigma_{12}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \sigma_{22}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \\
\Sigma^{-1} &= \begin{pmatrix} \frac{\sigma_{22}}{\Delta} & -\frac{\sigma_{12}}{\Delta} \\ -\frac{\sigma_{12}}{\Delta} & \frac{\sigma_{11}}{\Delta} \end{pmatrix}
\end{aligned}
$$

Put $\delta\mu_i = (\mu_i^A - \mu_i^B)$.

Then we have

$$
\begin{aligned}
D_\mu(f^A, f^B) &= \\
\sqrt{\delta\mu^T \cdot \Sigma^{-1} \cdot \delta\mu} &= \sqrt{\frac{\delta\mu_2\left(\sigma_{11}\delta\mu_2 - \sigma_{12}\delta\mu_1\right)}{\Delta} + \frac{\delta\mu_1\left(\sigma_{22}\delta\mu_1 - \sigma_{12}\delta\mu_2\right)}{\Delta}}.
\end{aligned}
$$

Numerical example:

$$
\Sigma^A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma^B = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}, \quad \Sigma^{B^{-1}} = \begin{pmatrix} 3/7 & -1/7 \\ -1/7 & 3/14 \end{pmatrix}
$$

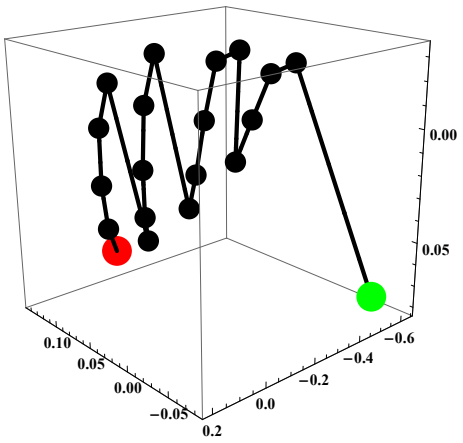

Figure 9: *Embedding of 20 evaluations of information distance for the bivariate covariances arising from a Poisson line process of line segments, (15), with* $x = \lambda = 1$ *and* $m = 2, 3, \ldots 21$. *The starting green point in the lower right is for* $m = 2$ *and the red end point is for* $m = 21$.

$$\Sigma^{A^{-1/2}} \cdot \Sigma^B \cdot \Sigma^{A^{-1/2}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix},$$

$$\text{with eigenvalues : } \lambda_1 = 7, \lambda_2 = 2.$$

$$D_\Sigma(\Sigma^A, \Sigma^B) = \sqrt{\frac{1}{2}\sum_{j=1}^{n} \log^2(\lambda_j)} \approx 1.46065$$

$$\Delta_\Sigma(\Sigma^A, \Sigma^B) = \sqrt{\log \frac{7+2}{4}} \approx 0.9005.$$

For comparison, the symmetrized Kullback-Leibler distance [16] is given by

$$D_{KL}(\Sigma^A, \Sigma^B) = \frac{1}{2}\left(\sqrt{\frac{1}{2}\log 14 - \frac{19}{28}} + \sqrt{\frac{1}{2}\log\frac{1}{14} + \frac{7}{2}}\right) \approx 1.1386.$$

6 Analysis of samples

6.1 Analytic results for Poisson processes of line segments and rectangles

We provide here some graphics showing three dimensional embeddings of Poisson processes that yield stochastic textures of pixel density, using

the analysis in §3.

Figure 9 shows an embedding of 20 samples calculated for a Poisson line process of line segments, (15), with $x = \lambda = 1$ and $m = 2, 3, \ldots 21$. The starting green point in the lower right is for $m = 2$ and the red end point is for $m = 21$. Figure 10 shows an embedding of 18 samples calculated for a planar Poisson process of unit squares, from (10), with $\omega = \lambda = 1$. It shows the separation into two groups of samples: analysed with small base pixels, $x = 0.1$ right, and with large base pixels, $x = 1$ left. Figure 11 shows an embedding of 18 samples calculated for a planar Poisson process of rectangles with aspect ratio 5:1, from (10), with $\omega = 0.2, \lambda = 1$. Again it shows the separation into two groups of samples analysed with small pixels, right, and with large pixels, left.

6.2 Deviations from Poisson arising from clustering

Our three spatial variables for each spatial array of data are the mean density in a central pixel, mean of its first neighbours, and mean of its second neighbours. We begin with analysis of a set of 16 samples of areal density maps for simulated stochastic fibre networks made from the same number of 1mm fibres but with differing scales (clump sizes) and intensities (clump densities) of fibre clustering. Among these is the standard unclustered Poisson fibre network; all samples have the same mean density.

Figure 12 gives analyses for spatial arrays of pixel density differences from Poisson networks. It shows a plot of $D_\Sigma(f^A, f^B)$ as a cubic-smoothed surface (left), and as a contour plot (right), for geodesic information distances among 16 datasets of 1mm pixel density differences between a Poisson network and simulated networks made from 1mm fibres. Each network has the same mean density but with different scales and densities of clustering; thus the mean difference is zero in this case. Second row: Dimensionality reduction embedding of the same data grouped by numbers of fibres in clusters and cluster densities. Using pixels of the order of fibre length is appropriate for extracting information on the sizes of typical clusters. The embedding reveals the clustering features as orthogonal subgroups.

Next, Figure 13 gives analyses for pixel density arrays of the clustered networks. It shows on the left the plot of $D_\Sigma(f^A, f^B)$ as a cubic-smoothed surface (left) for trivariate Gaussian information distances among the 16 datasets of 1mm pixel densities for simulated networks made from 1mm fibres, each network with the same mean density but with different clustering. In this case the trivariate Gaussians all have the same mean vectors. Shown on the right is the dimensionality re-

duction embedding of the same data grouped by numbers of fibres in clusters and cluster densities; the solitary point is a Poisson network of the same fibres.

6.3 Effect of mean density in Poisson structures

Figure 14 gives analyses for pixel density arrays for Poisson networks of different mean density. It shows the plot of $D_\Sigma(f^A, f^B)$ as a cubic-smoothed surface (left), for trivariate Gaussian information distances among 16 simulated Poisson networks made from 1mm fibres, with different mean density, using pixels at 1mm scale. Also shown is, (right) dimensionality reduction embedding of the same Poisson network data, showing the effect of mean network density.

6.4 Analysis of commercial samples

Figure 15 shows a 3-dimensional embedding for a data set from [9] including 182 paper samples from gap formers, handsheets, pilot machine samples and hybrid formers. We see that to differing degrees the embedding separates these different and very disparate forming methods by assembling them into subgroups. This kind of discrimination could be valuable in evaluating trials, comparing different installations of similar formers and for identifying anomalous behaviour.

The benefit from these analyses is the representation of the important structural features of number of fibres per cluster and cluster density, by almost orthogonal subgroups in the embedding.

6.5 Analysis of Saccharomyces Cerevisiae yeast and human genomes

This yeast is the genome studied in [3] for which we showed that all 20 amino acids along the protein chains exhibited mutual clustering, and separations of 3-12 are generally favoured between repeated amino acids, perhaps because this is the usual length of secondary structure, cf. also [1]. The database of sample sequences is available on the Saccharomyces Genome Database [19]. Here we mapped the sequences of the 20 amino acids A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y onto the 20 grey-level values $0.025, 0.075, \ldots, 0.975$ so yielding a grey-level barcode for each sequence, Figure 1. Given the usual length of secondary structure to range from 3 to 12 places along a sequence, we used spatial covariances between each pixel and its successive 12 neighbours. Figure 16 plots the determinants of the 12-variate spatial covariances of 20

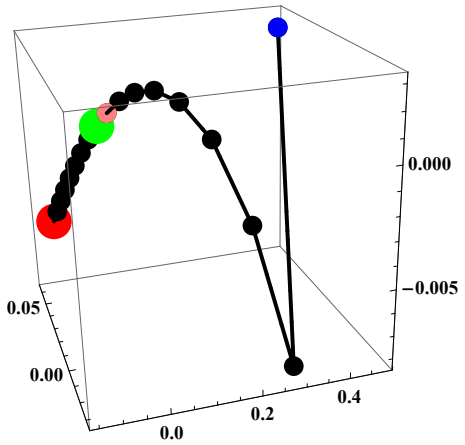

Figure 10: *Embedding of 18 evaluations of information distance for the bivariate covariances arising from a planar Poisson process of squares, (10), with* $\omega = \lambda = 1$. *The two groups arise from different schemes of inspection pixels. Right group used small base pixels with* $x = 0.1$, *from blue to pink* $m = 2, 3, \ldots, 10$; *Left group used large base pixels with* $x = 1$, *from green to red* $m = 2, 3, \ldots, 10$.

for yeast, black **Y**, together with 3 Poisson random sequences of 100,000 amino acids with the yeast relative abundances, blue **RY**. Also shown are 20 samples of human sequences, red **H**, and 3 Poisson sequences of 100,000 amino acids with the human relative abundances, green **RH**. Figure 17 shows an embedding of these 20 12-variate spatial covariances for yeast, small black points, together with 3 Poisson sequences of 100,000 amino acids with the yeast relative abundances, large blue points, and 20 human DNA sequences, medium red points using data from the NCBI Genbank Release 197.0 [15], and 3 Poisson sequences of 100,000 amino acids with the human relative abundances, large green points. The sequences ranged in length from 340 to 1900 amino acids. As with the original analysis of recurrence spacings [3] which revealed clustering, the difference of the yeast and human sequence structures from Poisson is evident. However, it is not particularly easy to distinguish yeast from human sequences by this technique, both lie in a convex region with the Poisson sequences just outside, but there is much scatter. Further analyses of genome structures will be reported elsewhere.

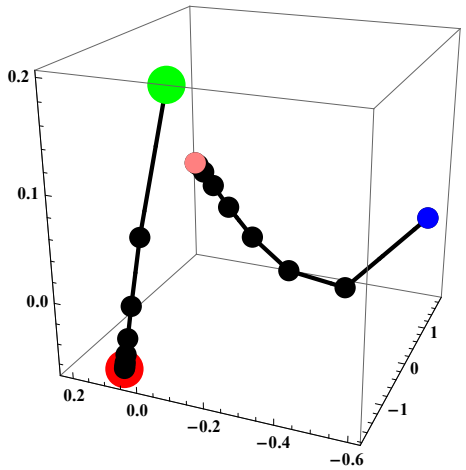

Figure 11: *Embedding of 22 evaluations of information distance for the bivariate covariances arising from a planar Poisson process of rectangles, (10), with* $\omega = 0.2, \lambda = 1$. *The two groups arise from different schemes of inspection pixels. Left group used large base pixels* $x = 1$, *from green to red* $m = 2, 3, \ldots, 10$; *Right group used small base pixels* $x = 0.1$, *from blue to pink* $m = 2, 3, \ldots, 10$.

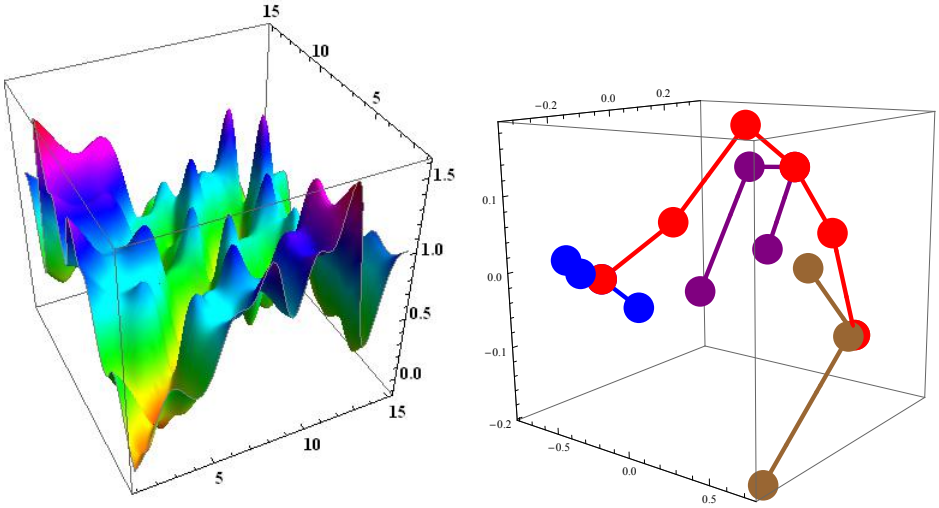

Figure 12: *Pixel density differences from Poisson networks. Left: plot of* $D_\Sigma(f^A, f^B)$ *as a cubic-smoothed surface, for trivariate Gaussian information distances among 16 datasets of 1mm pixel density differences between a Poisson network and simulated networks made from 1mm fibres, each network has the same mean density but with different clustering. Right: Embedding of the same data grouped by numbers of fibres in clusters and cluster densities.*

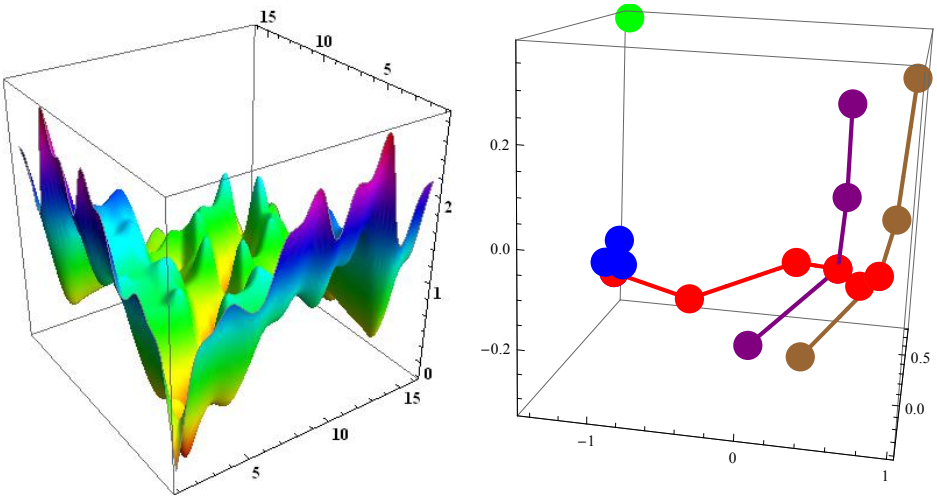

Figure 13: *Pixel density arrays for clustered networks: Left: plot of* $D_\Sigma(f^A, f^B)$ *as a cubic-smoothed surface, for trivariate Gaussian information distances among 16 datasets of 1mm pixel density arrays for simulated networks made from 1mm fibres, each network with the same mean density but with different clustering. Right: Embedding of the same data grouped by numbers of fibres in clusters and cluster densities; the solitary point is an unclustered Poisson network.*

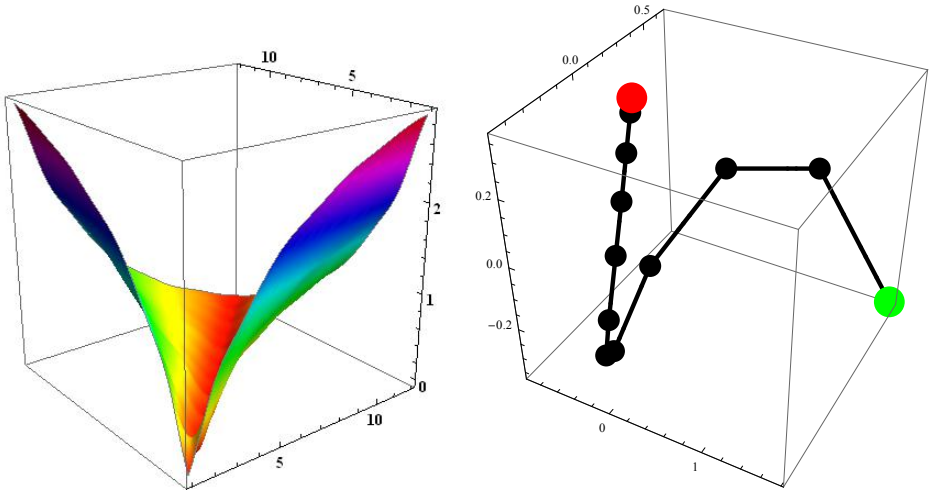

Figure 14: *Pixel density arrays for Poisson networks of different mean density. Left: Plot of* $D_\Sigma(f^A, f^B)$ *as a cubic-smoothed surface (left), for trivariate Gaussian information distances among 16 simulated Poisson networks made from 1mm fibres, with different mean density, using pixels at 1mm scale. Right: Embedding of the same Poisson network data, showing the effect of mean network density.*

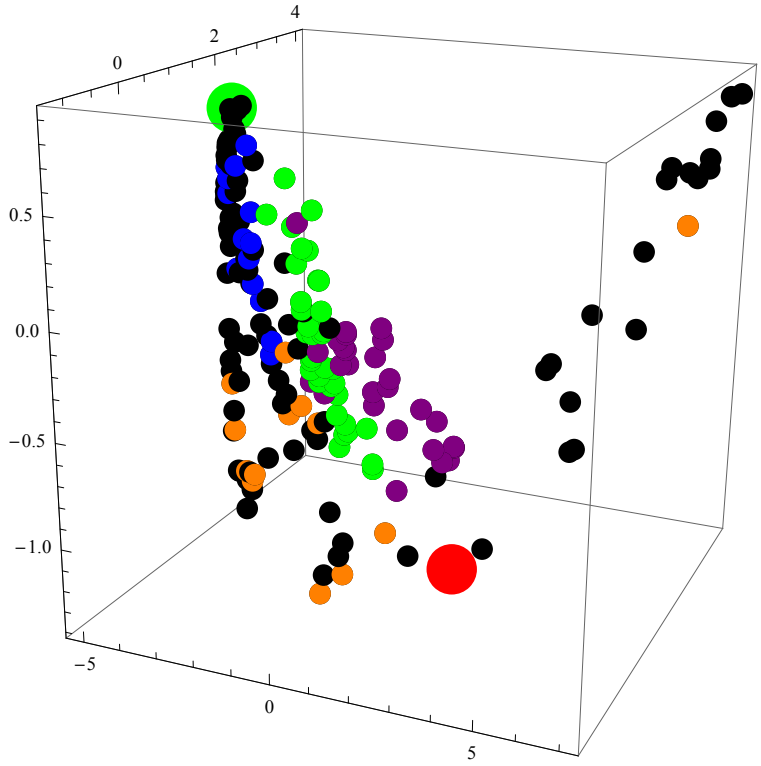

Figure 15: *Embedding using 182 trivariate Gaussian distributions for samples from the data set [9]. Blue points are from gap formers; orange are various handsheets, purple are from pilot paper machines and green are from hybrid formers. The embedding separates these different forming methods into subgroups.*

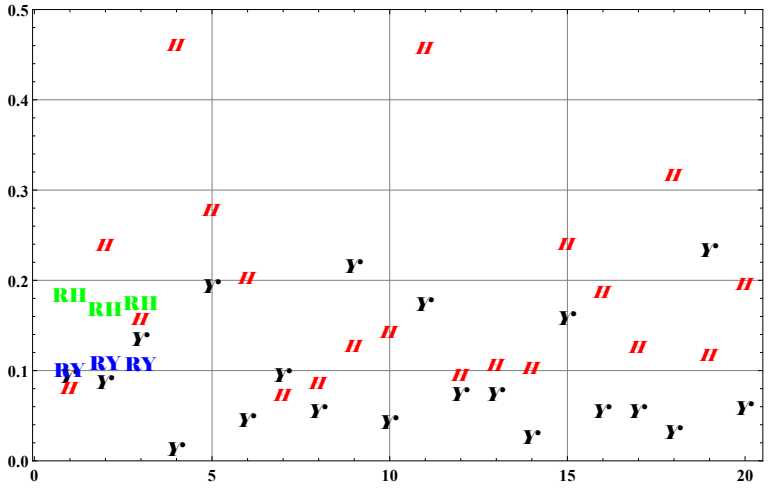

Figure 16: *Determinants of 12-variate spatial covariances for 20 samples of yeast amino acid sequences, black* **Y**, *together with 3 Poisson sequences of 100,000 amino acids with the yeast relative abundances, blue* **RY**. *Also shown are 20 samples of human sequences, red* **H**, *and 3 Poisson sequences of 100,000 amino acids with the human relative abundances, green* **RH**.

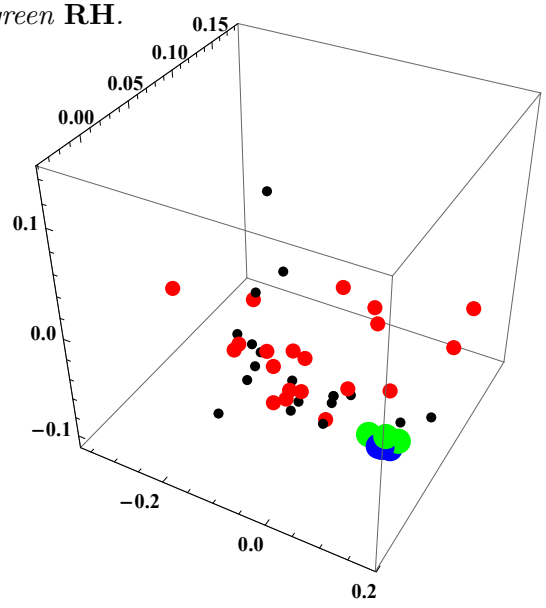

Figure 17: *12-variate spatial covariance embeddings for 20 samples of yeast amino acid sequences, small black points, together with 3 Poisson sequences of 100,000 amino acids with the yeast relative abundances, large blue points. Also shown are 20 human DNA sequences, medium red points, and 3 Poisson sequences of 100,000 amino acids with the human relative abundances, large green points.*

References

[1] K. Arwini and C.T.J. Dodson. **Information Geometry Near Randomness and Near Independence.** Lecture Notes in Mathematics. Springer-Verlag, New York, Berlin, 2008, Chapter 9 with W.W. Sampson, Stochasic Fibre Networks pp 161-194.

[2] C. Atkinson and A.F.S. Mitchell. Rao's distance measure. *Sankhya: Indian Journal of Statistics* 48, A, 3 (1981) 345-365.

[3] Cai, Y. Dodson, C.T.J. Wolkenhauer O. and Doig, A.J. (2002). *Gamma Distribution Analysis of Protein Sequences shows that Amino Acids Self Cluster.* Journal Theoretical Biology 218, 4 409-418.

[4] K.M. Carter, R. Raich and A.O. Hero. Learning on statistical manifolds for clustering and visualization. In 45th Allerton Conference on Communication, Control, and Computing, Monticello, Illinois, 2007. https://wiki.eecs.umich.edu/global/data/hero/images/c/c6/Kmcarter-learnstatman.pdf

[5] K.M. Carter **Dimensionality reduction on statistical manifolds**. PhD thesis, University of Michigan, 2009. http://tbayes.eecs.umich.edu/kmcarter/thesis

[6] M. Deng and C.T.J. Dodson. **Paper: An Engineered Stochastic Structure**. Tappi Press, Atlanta, 1994.

[7] C.T.J. Dodson. Spatial variability and the theory of sampling in random fibrous networks. *J. Roy. Statist. Soc. B* 33, 1 (1971) 88-94.

[8] C.T.J. Dodson. A geometrical representation for departures from randomness of the inter-galactic void probablity function. http://arxiv.org/abs/0811.4390 Workshop on Statistics of Cosmological Data Sets NATO-ASI Isaac Newton Institute 8-13 August 1999.

[9] C.T.J. Dodson, W.K. Ng and R.R. Singh. **Paper: Stochastic Structure Analysis Archive.** Pulp and Paper Centre, University of Toronto 1995, 3 CDs.

[10] C.T.J. Dodson and W.W. Sampson. In **Advances in Pulp and Paper Research, Oxford 2009**. *Trans. XIVth Fund. Res. Symp.* (S.J. I'Anson, ed.), pp665-691, FRC, Manchester, 2009.

[11] C.T.J. Dodson and W.W. Sampson. Dimensionality reduction for classification of stochastic fibre radiographs In Proc. **GSI2013 - Geometric Science of Information**, Paris, 28-30 August 2013. Lecture Notes in Computer Science 8085, Springer-Verlag, Berlin, 2013.

[12] A.G. Doroshkevich, D.L. Tucker, A. Oemler, R.P. Kirshner, H. Lin, S.A. Shectman, S.D. Landy and R. Fong. Large- and Superlarge-scale Structure in the Las Campanas Redshift Survey. *Mon. Not. R. Astr. Soc.* 283 4 (1996) 1281-1310.

[13] B. Ghosh. Random distances within a rectangle and between two rectangles. *Calcutta Math. Soc.* 43(1) (1951) 17-24.

[14] K.V. Mardia, J.T. Kent and J.M. Bibby. **Multivariate Analysis** Academic Press, 1980.

[15] NCBI Genbank of The National Center for Biotechnology Information. Samples from CCDS_protein.20130430.faa.gz ftp://ftp.ncbi.nlm.nih.gov/genbank/README.genbank

[16] F. Nielsen, V. Garcia and R. Nock. Simplifying Gaussian mixture models via entropic quantization. In Proc. 17^{th} European Signal Processing Conference, Glasgow, Scotland 24-28 August 2009, pp 2012-2016.

[17] W.W. Sampson. **Modelling Stochastic Fibre Materials with *Mathematica*.** Springer-Verlag, New York, Berlin, 2009.

[18] W.W. Sampson. Spatial variability of void structure in thin stochastic fibrous materials. *Mod. Sim. Mater. Sci. Eng.* 20:015008, (2012), pp13. doi:10.1088/0965-0393/20/1/015008

[19] Saccharomyces Cerevisiae yeast Genome Database `http://downloads.yeastgenome.org/sequence/S288C_reference/orf_protein/`