

*GSI2013 - Geometric Science of Information,
Paris, 28-30 August 2013*

Dimensionality reduction for classification of stochastic fibre radiographs

C.T.J. Dodson¹ and W.W. Sampson²

School of Mathematics¹ and School of Materials²
University of Manchester UK

ctdodson@manchester.ac.uk

Abstract

Dimensionality reduction helps to identify small numbers of essential features of stochastic fibre networks for classification of image pixel density datasets from experimental radiographic measurements of commercial samples and simulations.

Typical commercial macro-fibre networks use finite length fibres suspended in a fluid from which they are continuously deposited onto a moving bed to make a continuous web; the fibres can cluster to differing degrees, primarily depending on the fluid turbulence, fibre dimensions and flexibility.

Here we use information geometry of trivariate Gaussian spatial distributions of pixel density among first and second neighbours to reveal features related to sizes and density of fibre clusters.

Introduction

Much analytic work has been done on modelling of the statistical geometry of stochastic fibre networks and their behaviour in regard to strength, fluid ingress or transfer [1, 5, 7].

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. Clustering of fibres is well-approximated by Poisson processes of Poisson clusters of differing density and size.

A Poisson fibre network is a standard reference structure for any given size distribution of fibres; its statistical geometry is well-understood for finite and infinite fibres.

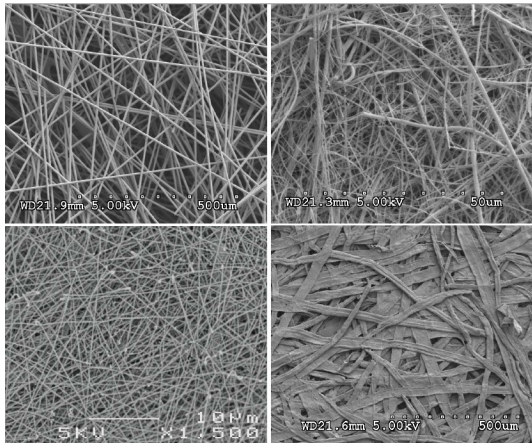


Figure : 1. *Electron micrographs of four stochastic fibrous materials. Top left: Nonwoven carbon fibre mat; Top right: glass fibre filter; Bottom left: electrospun nylon nanofibrous network (Courtesy S.J. Eichhorn and D.J. Scurr); Bottom right: paper using wood cellulose fibres—typically flat ribbonlike,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

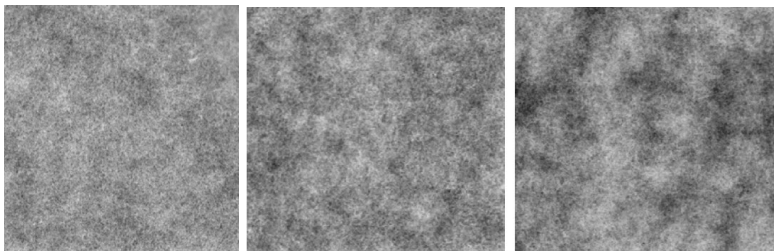


Figure : 2. *Areal density radiographs of three paper networks made from natural wood cellulose fibres, of order 1 mm in length, with constant mean density but different distributions of fibres. Each image represents a square region of side length 5 cm; darker regions correspond to higher coverage. The left image is similar to that expected for a Poisson process of the same fibres, so typical real samples exhibit clustering of fibres.*

Spatial statistics

We 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 fibre clusters, which could arise in one, two or three dimensions—the graphic shows a grey level barcode for the ordered sequence of the 20 amino acids in a yeast genome, a 1-dimensional stochastic texture.

Saccharomyces Cerevisiae Amino Acids SC1



For isotropic spatial processes, which we consider here, the variables are means over shells of first and second neighbours, respectively, which share the population mean with the central pixel. For anisotropic networks the neighbour groups would be split into more, orthogonal, new variables to pick up the spatial anisotropy in the available spatial directions.

Typical sample data

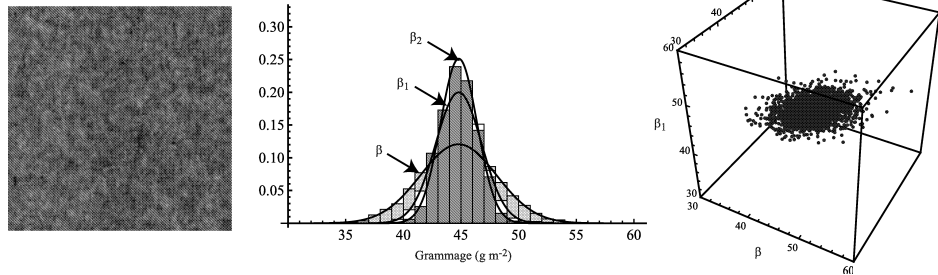


Figure : 3. Trivariate distribution of areal density values for a typical newsprint sample. Left: source radiograph; centre: histogram of pixel densities $\tilde{\beta}_i$, average of first neighbours $\tilde{\beta}_{1,i}$ and second neighbours $\tilde{\beta}_{2,i}$; right: 3D scatter plot of $\tilde{\beta}_i$, $\tilde{\beta}_{1,i}$ and $\tilde{\beta}_{2,i}$.

Information geodesic distances between multivariate Gaussians

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

$D_\mu(f^A, f^B)$ when they have a common mean μ but different covariances Σ^A, Σ^B and

$D_\Sigma(f^A, f^B)$ when they have a common covariance Σ but different means μ^A, μ^B .

The general case is not known analytically but for the purposes of studying the stochastic textures arising from areal density arrays of samples of stochastic fibre networks, a satisfactorily discriminating approximation is

$$D(f^A, f^B) \approx D_\mu(f^A, f^B) + D_\Sigma(f^A, f^B).$$

Information geodesic distance between multivariate Gaussians [2]

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

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

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

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

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

From the form of $D_{\Sigma}(f^A, f^B)$ in (2) it may be seen that an approximate monotonic relationship arises with a more easily computed symmetrized log-trace function given by $\Delta_{\Sigma}(f^A, f^B) =$

$$\sqrt{\log \left(\frac{1}{2n} \left(\text{Tr}(\Sigma^{A-1/2} \cdot \Sigma^B \cdot \Sigma^{A-1/2}) + \text{Tr}(\Sigma^{B-1/2} \cdot \Sigma^A \cdot \Sigma^{B-1/2}) \right) \right)}. \quad (3)$$

This is illustrated by the plot of $D_{\Sigma}(f^A, f^B)$ from equation (2) on $\Delta_{\Sigma}(f^A, f^B)$ from equation (3) in Figure 4 for 185 trivariate Gaussian covariance matrices.

For comparing relative proximity, this is a better measure near zero than the symmetrized Kullback-Leibler distance [6] in those multivariate Gaussian cases so far tested and may be quicker for handling large batch processes.

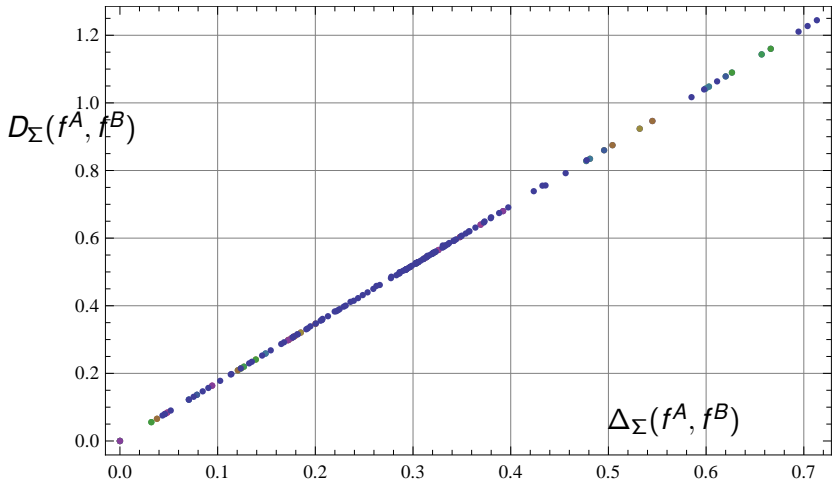


Figure : 4. Plot of $D_{\Sigma}(f^A, f^B)$ from (2) on $\Delta_{\Sigma}(f^A, f^B)$ from (3) for 185 trivariate Gaussian covariance matrices.

Dimensionality reduction for data sets

1. Obtain mutual 'information distances' $D(i, j)$ among the members of the data set of textures X_1, X_2, \dots, X_N each with 250×250 pixel density values.
2. The array of $N \times N$ differences $D(i, j)$ is a symmetric positive definite 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 symmetric positive definite with diagonal zero. We compute its N eigenvalues $ECD(i)$, which are necessarily real, and find 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.

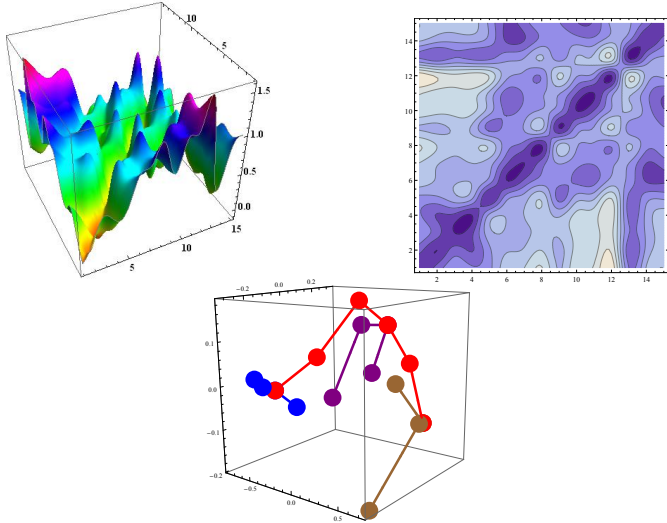


Figure : 5. $D_{\Sigma}(f^A, f^B)$ as a cubic-smoothed surface (left), contour plot (right), trivariate Gaussian information distances among 16 datasets of 1mm pixel density differences between a Poisson network and simulated networks from 1mm fibres, same mean density different clustering. Embedding: subgroups show numbers of fibres in clusters and cluster densities.

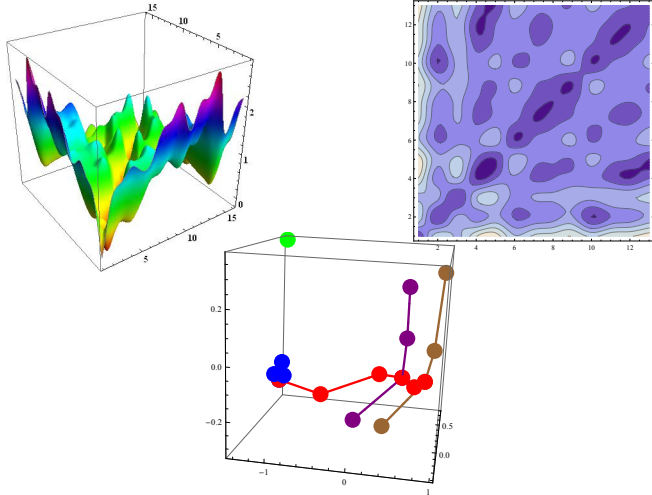


Figure : 6. $D_{\Sigma}(f^A, f^B)$ as a cubic-smoothed surface (left), contour plot (right), for trivariate Gaussian information distances among 16 datasets of 1mm pixel density arrays for simulated networks made from 1mm fibres, each network with the same mean density but with different clustering. Embedding: subgroups show numbers of fibres in clusters and cluster densities; the solitary point is an unclustered Poisson network.

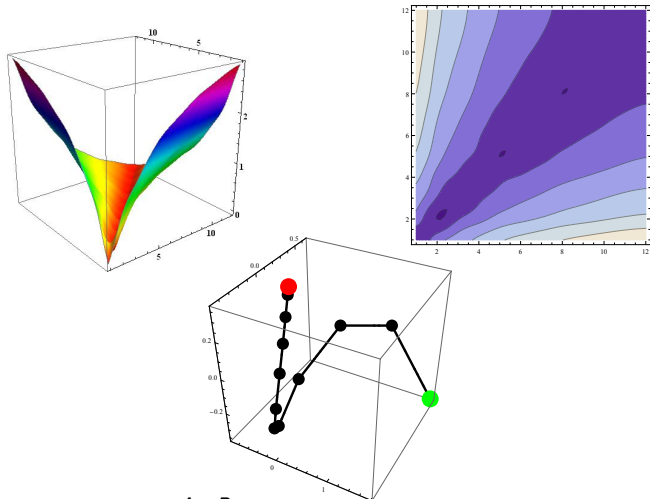


Figure : 7. $D_{\Sigma}(f^A, f^B)$ as a cubic-smoothed surface (left), and as a contour plot (right), for trivariate Gaussian information distances among 16 simulated Poisson networks made from 1mm fibres, with different mean density, using pixels at 1mm scale. Second row: Embedding of the same Poisson network data, showing the effect of mean network density.

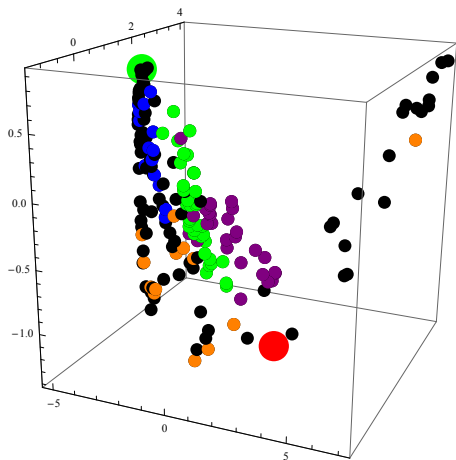







Figure : 8. *Embedding using 182 trivariate Gaussian distributions for samples from a data set of radiographs of commercial papers. The embedding separates different forming methods into subgroups.*

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, Stochastic 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] 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>
-  [4] K.M. Carter **Dimensionality reduction on statistical manifolds**. PhD thesis, University of Michigan, 2009.
<http://tbayes.eecs.umich.edu/kmcarter/thesis>
-  [5] M. Deng and C.T.J. Dodson. **Paper: An Engineered Stochastic Structure**. Tappi Press, Atlanta, 1994.



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



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