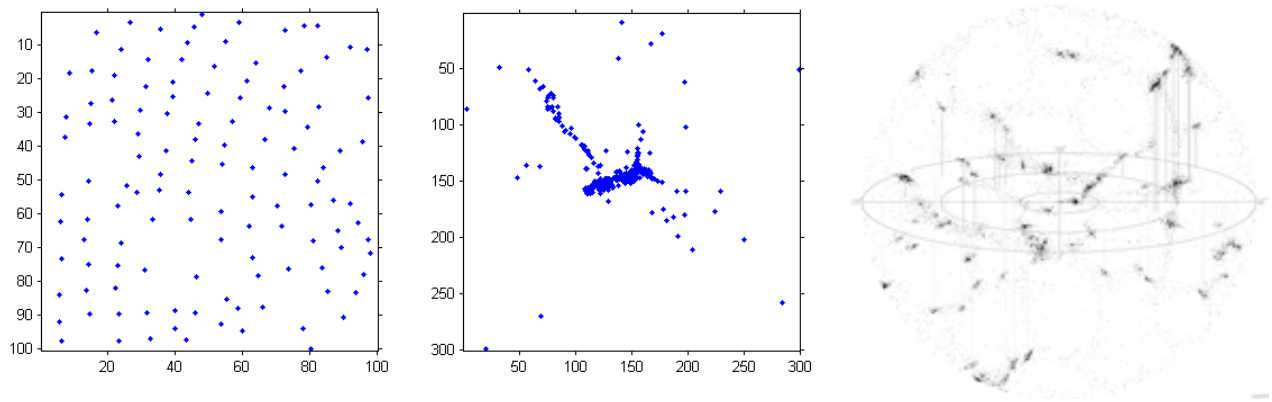
# A Gradient Field Approach to Modelling Fibre-Generated Spatial Point Processes
# (La modélisation des processus ponctuels spatiaux générés par les fibres: une méthode de champ de gradient)

Hill, Bryony
*University of Warwick, Department of Statistics*
*Coventry CV4 7AL, UK*
*E-mail:* B.J.Hill@warwick.ac.uk

Kendall, Wilfrid S.
*University of Warwick, Department of Statistics*
*Coventry CV4 7AL, UK*
*E-mail:* W.S.Kendall@warwick.ac.uk

Thönnes, Elke
*University of Warwick, Department of Statistics*
*Coventry CV4 7AL, UK*
*E-mail:* E.Thonnes@warwick.ac.uk



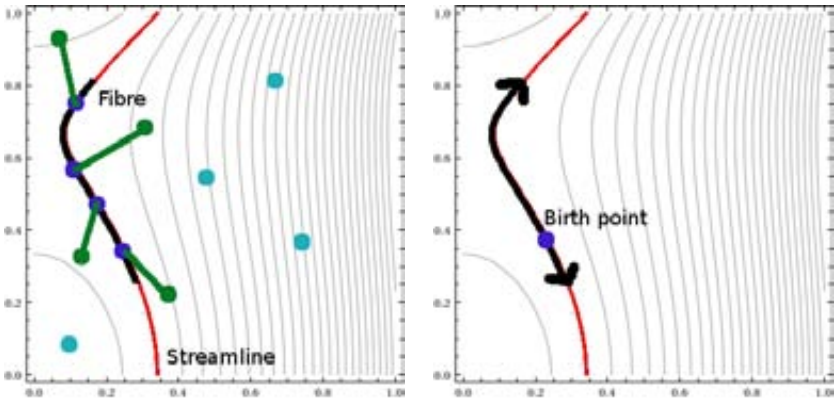

(a) Fingerprint sweat pores. (b) Earthquake epicentres. (c) Universe within 500 Mly of Earth.

Figure 1: (a) Fingerprint data extracted from a portion of fingerprint a002-5 from NIST Special Database 30 (Watson, 2001). (b) Earthquake data: epicentres in New Madrid region, taken from CERI (Center for Earthquake Research and Information). (c) Universe image: Richard Powell (`atlasoftheuniverse.com/nearsc.html`: Creative Commons Attribution-ShareAlike 2.5 License).

This paper is a preliminary report on a particular problem concerning inference of unobserved curvilinear structure. The problem is as follows: given an observed pattern of points in the plane or in space, draw reasonable conclusions concerning unobserved curvilinear fibres along which the points are supposed to cluster. Examples can occur at strikingly different length-scales as illustrated in Figure 1: (a) the point pattern of sweat pores lying along ridges as observed in a fingerprint; (b) the pattern of historical earthquake locations corresponding to unobserved geological faults; (c) the spatial locations of galaxies thought to cluster in unobserved filamentary structures. Various approaches to the inferential question have been proposed; we mention the method of principal curves (Hastie and Stuetzle, 1989; see in particular the algorithm proposed in Stanford and Raftery, 2000), the approach of Stoica, Martínez and Saar (2007) using Candy and Bisous models, and a precursor to the current

paper which introduced non-standard clustering ideas related to Diffusion Tensor Imaging (Su *et al*, 2008; Su, 2009). The work described briefly here is being written up for the PhD thesis of the first author (Hill, 2011), and will also be expounded in detail and at greater length in a forthcoming paper.

Here we report on work which uses the clustering approach of Su *et al* (2008) and Su (2009) to build an Empirical Bayes approach to the inferential problem. In the first place we suppose that the 2-d or 3-d window of observation is decomposed into a (random) *fibration* (a disjoint union of a whole continuum family of one-dimensional smooth curves – think of streamlines in a fluid). As illustrated in Figure 2, a finite random collection of continuous segments of these curves or streamlines is used to model the unobserved fibres generating the latent curvilinear structure of the observed point pattern. For each of the fibres, a random choice is made of a sequence of *mark points* along the fibre; the observed point set is formed by the union of random perturbations of these mark points (these perturbations forming the set of *signal points*) together with a further random scattering of points unattached to fibres (the set of *noise points*).




(a) Fibration: streamline and fibre. (b) Fibration: construction of fibre.

Figure 2: (a) Filtration with specified streamline and fibre. Mark points are indicated along the fibre and used to generate perturbed signal points. Other noise points are scattered over the entire window. (b) Construction of fibre using birth point together with random lengths in both directions of the fibre.

Some care is required if we are to make good mathematical and statistical sense of this construction. Given the random fibration by curves ("streamlines"), one might model the set of fibres by a random selection of a finite number of connected fragments obtained by breaking each of the streamlines into pieces using independent Poisson point processes along the streamlines. However that would entail the mathematically precarious construction of an uncountable number of independent Poisson point processes (one for each streamline); a hazardous probabilistic activity needing extreme measure-theoretic care. So we suppose instead that the fragments are formed by picking a finite number of "birth points" from the observation window; and, for each such birth point, proceeding in either direction of the fibre by random lengths so as to generate a fibre as a connected fragment of the streamline. (We make no attempt to control biases that might be introduced by this procedure – indeed, it would require a careful treatment to make consideration even of what "bias" might mean in this context; we plan to return to this matter at a later date.) The advantage of the "birth point" construction is that it is easily represented as the equilibrium state of a reversible Markov chain Monte Carlo algorithm, in which fibres die at random, and are born according to the construction above.

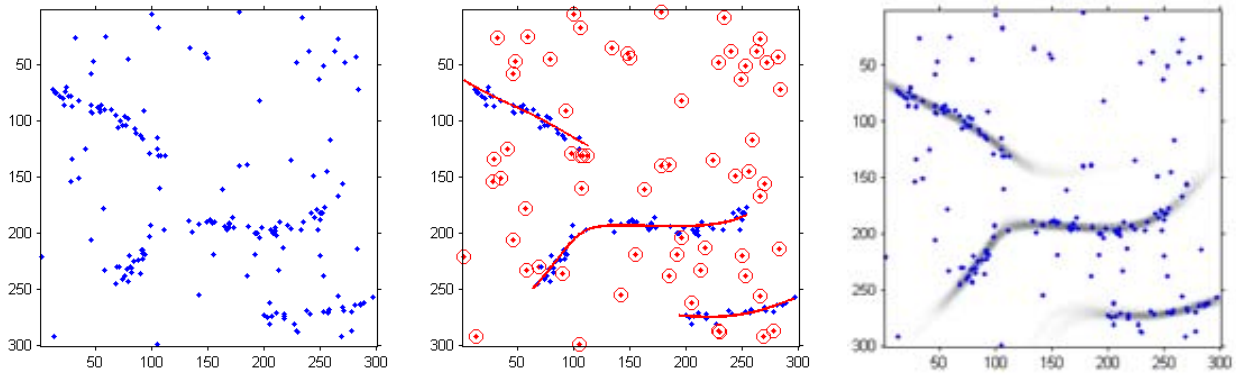
Given a fibre, we form a stationary random sequence of mark points along the fibre. The most direct way is as described above (use one-dimensional Poisson point processes); however we found it preferable to use the more general class of stationary renewal processes with Gamma-distributed inter-point distances, as this allows us to model a tendency to regularity along the fibre. Finally the actual signal points are modelled as Gaussian perturbations of the mark points, while the noise points

are modelled by an independent spatial Poisson point process.

Conditional on the fibration, and extending our remark concerning the construction of the fibres, the entire construction may be represented as the equilibrium state of a reversible Markov chain Monte Carlo algorithm. Moreover this construction may be augmented to allow Gibbs' updates of various parameters, to introduce latent variables expressing the probabilities of specified observed points being signal or noise, and to improve mixing by allowing movement and splitting and joining of fibres (implemented by manipulating the fibre birth points and upstream/downstream lengths). Finally, it is possible to run the algorithm without allowing the observed point locations to alter, thus simulating the conditional distribution of fibres and other parameters. *However* this conditional distribution is in fact conditional on the fibration as well as on the observed point locations, and we need to deal with this extra conditioning in order to produce a useful inferential procedure.

In principle one might address this issue by positing a prior distribution on the space of all fibrations. However this forms a challenging prospect, especially as one would also need to produce a reversible Markov chain Monte Carlo algorithm which possessed this distribution as a stationary state and exhibited good mixing under conditioning. Therefore we resort to an empirical Bayes procedure; we estimate the fibration based on the pattern of observed points. We do this in the manner described by Su *et al* (2008) and Su (2009), by creating a local orientation at each observed point $x$. This is done as follows. Let $(r, \theta)$ be the location of an observed point in polar coordinates using $x$ as origin. Map $(r, \theta)$ to points $(\pm \exp(-r^2/\sigma^2), \theta)$ (for a given scale parameter $\sigma$); calculate the $(2 \times 2)$ inertia tensor for the set of transformed observed points, weighted by the corresponding probabilities for those points to be signal rather than noise, and use the tensor's principal eigenvector to generate the orientation of the fibration at that point. The orientation field of the fibration can then be viewed as a gradient field, and integrated to produce the fibration.

Two points should be noted here. Firstly, we smooth the resulting tensor field and interpolate it over the entire observation window by using tensor means with local weighting factors (a statistically motivated introduction to tensor means can be found in Dryden *et al*, 2009). Secondly, computation of the tensor mean additionally weights each observed point according to the probability of it being signal rather than noise; thus computation of both the tensor means and the inertial tensors themselves will change every time these probabilities change. It follows that the fibration itself changes when changes occur in the observed point probabilities of noise versus signal. This mitigates to some extent against the circular nature of the empirical Bayesian treatment of the prior for the fibration.




(a) Test data set. (b) Reconstruction (noise encircled). (c) Reconstruction (fibre intensity).

Figure 3: Illustration of results obtained for a test data set.

We will report more fully on the results achieved by this algorithm in the forthcoming paper; here we note that satisfactory results can be achieved if care is taken to make good choices of hyperparameters. To illustrate what can be achieved, we present the reconstruction of a simple test example

(Figure 3). Image (a) illustrates the unprocessed dataset; image (b) presents a typical reconstruction, in which points classified as noise are circled, and modelled fibres are included; while image (c) indicates the empirical intensity of the random set composed of the modelled fibres, averaged over a long Markov chain Monte Carlo sequence. A major advantage of our approach is that we can derive statistical summaries including highest posterior probability density intervals for measures of interest (number of fibres, total length of fibres, ...). Figure 3(c) illustrates this graphically by indicating the length intensity for fibre estimates. Details of reconstructions of real datasets, including statistical summaries, will be found in our forthcoming paper.

## REFERENCES

Dryden IL, Koloydenko A, Zhou D (2009), "Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging". *Annals of Applied Statistics*;3(3):1102–1123.

Hastie T, Stuetzle W (1989), "Principal curves". *Journal of the American Statistical Association*;84(406):502–516.

Stanford DC, Raftery AE (2000), "Finding curvilinear features in spatial point patterns: principal curve clustering with noise". *IEEE Transactions on Pattern Analysis and Machine Intelligence*;22(6):601–609.

Stoica RS, Martínez VJ, Saar E (2007), "A three-dimensional object point process for detection of cosmic filaments". *Journal of the Royal Statistical Society: Series C (Applied Statistics)*;56(4):459–477.

Su J (2009), *A Tensor Approach to Fingerprint Analysis*, PhD Thesis, University of Warwick.

Su J, Hill B, Kendall WS, Thönnes E (2008), "Inference for point processes with unobserved, one-dimensional reference structure". University of Warwick Department of Statistics *CRiSM Working Paper* No.8–10.

Watson, C (2001), NIST Special Database 30: Dual Resolution Images from Paired Fingerprint Cards. National Institute of Standards and Technology, Gaithersburg, Md, USA. 2001.

## RÉSUMÉ

*On introduit un nouveau modèle bruité non paramétrique pour les processus ponctuels, qui se rassemblent autour des courbes ou des fibres. Le modèle identifie les courbes aléatoires comme étant les lignes intégrales d'un champ de gradient. En principe, ceci permet l'inclusion de toute courbe qui ne s'intersecte pas, avec seule une contrainte de continuité sous-jacente. On mélange une méthode de Monte-Carlo avec une procédure empirique de Bayes, afin de fournir une procédure pratique d'estimation des propriétés de la distribution sous-jacente des fibres, procédure basée sur les motifs observés dans les données ponctuelles. On fait des comparaisons avec les différentes techniques de la littérature. Finalement, on illustre la mthodologie par des applications aux empreintes digitales, aux tremblements de terre et aux galaxies.*

## ABSTRACT

*A new non-parametric model is introduced for point processes that are clustered along curves or fibres, with additional background noise. The model identifies random curves as integral lines of a gradient field. In principle this enables the inclusion of all possible non-self-intersecting curves with one underlying smoothness constraint. Markov chain Monte Carlo is combined with Empirical Bayes to provide a practical estimation procedure for properties of the underlying fibre distribution, based on the observed point pattern data. Comparisons are made with other techniques in the literature. Illustrations of the methodology include applications to fingerprints, earthquakes and galaxies.*