Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

# Does Bayes beat squinting? Estimating unobserved aspects of a spatial cluster process

Note no SAMBA/05/18
Authors Alisha Albert-Green
Peter Guttorp
Thordis Thorarinsdottir
Date 28th September 2018

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

## The authors

Alisha Albert-Green is Postdoctoral Fellow at the University of Waterloo in Canada. Peter Guttorp is Professor II and Thordis L. Thorarinsdottir is Research Leader at the Norwegian Computing Center.

## Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

**Norsk Regnesentral**
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

**Besøksadresse**
*Office address*
Gaustadalléen 23
NO-0373 Oslo, Norway

**Telefon** · telephone
(+47) 22 85 25 00
**Telefaks** · telefax
(+47) 22 69 76 60

**Internett** · internet
www.nr.no
**E-post** · e-mail
nr@nr.no

| | |
|---|---|
| **Title** | **Does Bayes beat squinting? Estimating unobserved aspects of a spatial cluster process** |
| **Authors** | **Alisha Albert-Green, Peter Guttorp, Thordis Thorarinsdottir** |
| Date | 28th September 2018 |
| Publication number | SAMBA/05/18 |


## Abstract

A point process data set on epidermal nerve fiber bundles is used as the basis for a series of experiments in identifying clusters. In this data set we know which secondary points are connected to which primary points. We will pretend that we do not have this information, and using Bayesian tools estimate the information from data. For comparison we also use k-means clustering. We do this both for known cluster centers, and when the cluster centers must be estimated from data.



| | |
|---|---|
| Keywords | Cluster point process; Latent point process; Point process inference |
| Target group | Statisticians |
| Availability | Open |
| Project | PointProcess |
| Project number | 220708 |
| Research field | Statistics |
| Number of pages | 18 |

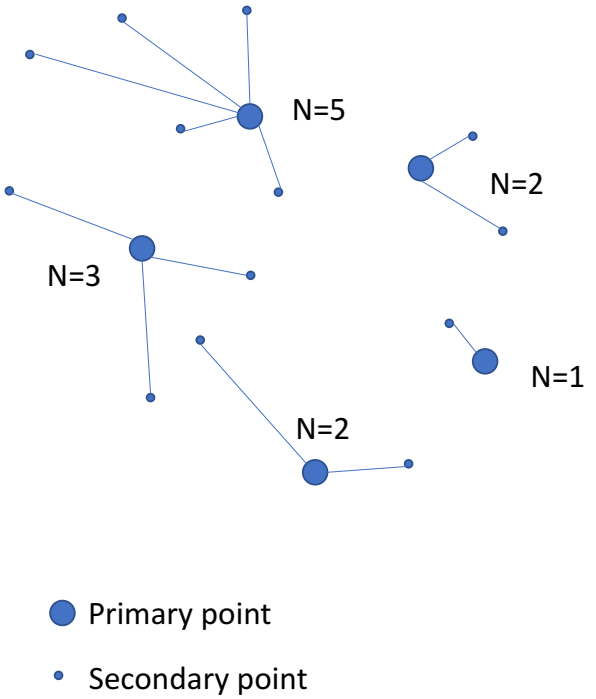


Figure 1. Schematic representation of Neyman-Scott cluster process.

# 1 Introduction

In the old days, finding a cluster of galaxies was apparently done by putting up a photograph of the heavens on the wall, stepping back from it, and squinting to determine which belong together (J. Neyman, personal communication to the second author). Neyman et al. (1956) outline a stochastic model for cluster of galaxies, specifying the assumptions needed for the model and outlining how to use data to estimate parameters in this well-known Neyman-Scott cluster model (Guttorp, 1995, p. 232). Figure 1 is a schematic representation of the process, where large dots correspond to cluster centers (primary points), lines connect secondary points to cluster centers, snd for each cluster the number $N$ of points is indicated. The stochastic structure is that the secondary points connected to a primary point are independent processes, obtained by putting a random number of points independently around the primary point.

In this work we will consider a different situation in which clusters also occur naturally (Waller et al., 2011), namely epidermal nerve fiber bundles. The data set contains both the common origins deep in the epidermis and the nerve entries near the top of the skin. Thus, we know the primary cluster centers, the secondary cluster points, the distances between primary and secondary points, and the number of points in each cluster. We will use a model similar to that developed by Andersson et al. (2016). Figure 2 shows the observations for one of the subjects. But our goal in this work is to investigate to what extent we can reconstruct unobserved aspects of a cluster process from observed parts and the model. Suppose for example that we know what points are primary and what

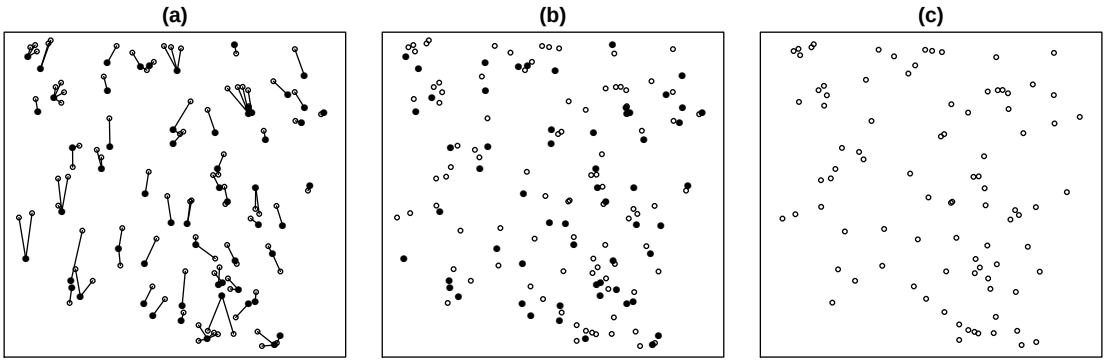


Figure 2. Observations for one individual: (a) Full cluster information with primary points indicated by black dots and secondary points indicated by circles, (b) primary and secondary points without family relations, (c) secondary points only.

points are secondary, but not which secondary points correspond to a given primary point. Can we successfully reconstruct the complete connectivity pattern?

Albert-Green (2016) developed a hierarchical cluster process for modelling spatio-temporal storm cell data in which a Neyman-Scott process, with the parents assumed to follow a log-Gaussian Cox process, was specified. Parameter estimation was accomplished by employing minimum contrast estimation for the spatial and temporal projection processes (Møller and Ghorbani, 2012; Prokešová and Dvořák, 2014). For both projections, the parameters in the second level of the hierarchy were estimated conditional on those of the first, as developed in Wiegand et al. (2007). Although these results were satisfactory, this required introducing extra tuning parameters. By extending the methodology developed in this paper, parameter estimation could be performed using a more parsimonious parametric approach.

We will, in subsequent sections, look at two different reconstruction problems, and using data from Kennedy et al. (1996) discuss algorithms for the reconstructions. After introducing some notation in section 2, we try to identify what secondary points belong to what cluster (knowing the cluster centers) in section 3, while in section 4 we try to determine the cluster centers observing only the secondary points. Some issues of discussion can be found in section 5.

# 2 Point process model with clustering

Write the primary process $\Psi = (\mathbf{c}, \gamma) = \{(c_j, \gamma)\}$ where the primary points $c_j \in \mathbb{R}^2$ and the expected cluster size $\gamma \in \mathbb{R}_+$. We will assume that the cluster centers follow a Poisson process. Following Baddeley (2010), we denote the offspring (secondary) process by $\Phi = (\mathbf{x}, \mathbf{m}) = \{(x_i, m_i)\}$, where the secondary points $x_i \in \mathbb{R}^2$ and the label (primary point indicator) $m_i \in \mathbb{Z}_+$. Note that conditional on $\Psi = \psi$, $m_i$ can only take values in $\{1, 2, ..., M\}$ where $M = n(\psi)$ is the number of cluster centers. We get the cluster size for

the cluster center $c_j$ by calculating $\sum_{i=1}^{N} \mathbb{1}\{m_i = j\}$, where $N = n(\phi)$ denotes the number of secondary points.

We assume the model for $\Phi$ is that of a doubly stochastic Poisson point process with the stochastic intensity given by

$$Z\left((x_i, m_i) \mid \psi, \boldsymbol{\theta}\right) = \gamma k(x_i, c_{m_i} \mid \boldsymbol{\theta}), \tag{1}$$

where $k$ is a kernel density describing the dispersion of cluster points around cluster centers, with identical parameters for all clusters. We define $k$ as a function of the squared distance $r^2 = \|x - c_m\|^2$ and the angle $u = \angle(x, c_m)$,

$$k(x, c_m \mid \boldsymbol{\theta}) = h(r^2 \mid \theta_1) g(u \mid \boldsymbol{\theta}_{-1}). \tag{2}$$

Here, $\boldsymbol{\theta}_{-1}$ denotes the parameter vector $\boldsymbol{\theta}$ without the first element $\theta_1$. The function $h$ is the density function of the exponential distribution,

$$h(r^2 \mid \theta_1) = \theta_1 e^{-\theta_1 r^2}, \tag{3}$$

with $\theta_1 > 0$. The angle density $g$ is given by the density of a mixture of two von Mises distributions,

$$g(u \mid \boldsymbol{\theta}_{-1}) = \theta_2 \frac{e^{\theta_4 \cos(u - \theta_3)}}{2\pi I_0(\theta_4)} + (1 - \theta_2) \frac{e^{\theta_6 \cos(u - \theta_5)}}{2\pi I_0(\theta_6)}, \tag{4}$$

where $I_0$ is the modified Bessel function of order 0, $\theta_2 \in [0, 1]$, $\theta_3, \theta_5 \in$ and $\theta_4, \theta_6 > 0$.

Without loss of generality, assume the observation window is $B = [0, 1]^2$ and define

$$A(\boldsymbol{\theta}) := \int_B \sum_{m=1}^{M} \gamma k(\xi, c_m \,|\, \boldsymbol{\theta}) \mathrm{d}\xi = \sum_{m=1}^{M} \gamma \int_B k(\xi, c_m \mid \boldsymbol{\theta}) d\xi.$$

Then, the conditional likelihood of the secondary process on $B$ can be written as

$$p\big((\mathbf{x}, \mathbf{m}) \mid \psi, \boldsymbol{\theta}\big) = \exp\Big(M - A(\boldsymbol{\theta})\Big) \gamma^N \prod_{i=1}^{N} k(x_i, c_{m_i} \mid \boldsymbol{\theta}). \tag{5}$$

The log-likelihood is then given by

$$\begin{aligned} &\log p\big((\mathbf{x},\mathbf{m}) \,|\, \psi, \boldsymbol{\theta}\big) \\ &\qquad = M - A^*(\boldsymbol{\theta}) + N \log \gamma + N \log \theta_1 - \theta_1 \sum_{i=1}^{N} r_i^2 + \sum_{i=1}^{N} \log g(u_i \mid \boldsymbol{\theta}_{-1}), \end{aligned} \tag{6}$$

where

$$A^*(\boldsymbol{\theta}) = \gamma \sum_{m=1}^{M} \int_B k(\xi, c_m \mid \boldsymbol{\theta}) d\xi.$$

From Baddeley (2010, Lemma 3) it follows that $\Phi \mid \Psi, \theta$ is a marked Poisson process. Hence the marginal distribution of the secondary point locations is Poisson with rate

$$Z_m\left(\xi \mid \psi, \boldsymbol{\theta}\right) = \sum_{m=1}^{n(\psi)} \gamma k(\xi, c_m \mid \boldsymbol{\theta}). \tag{7}$$

# 3 Simulation of the Point Process

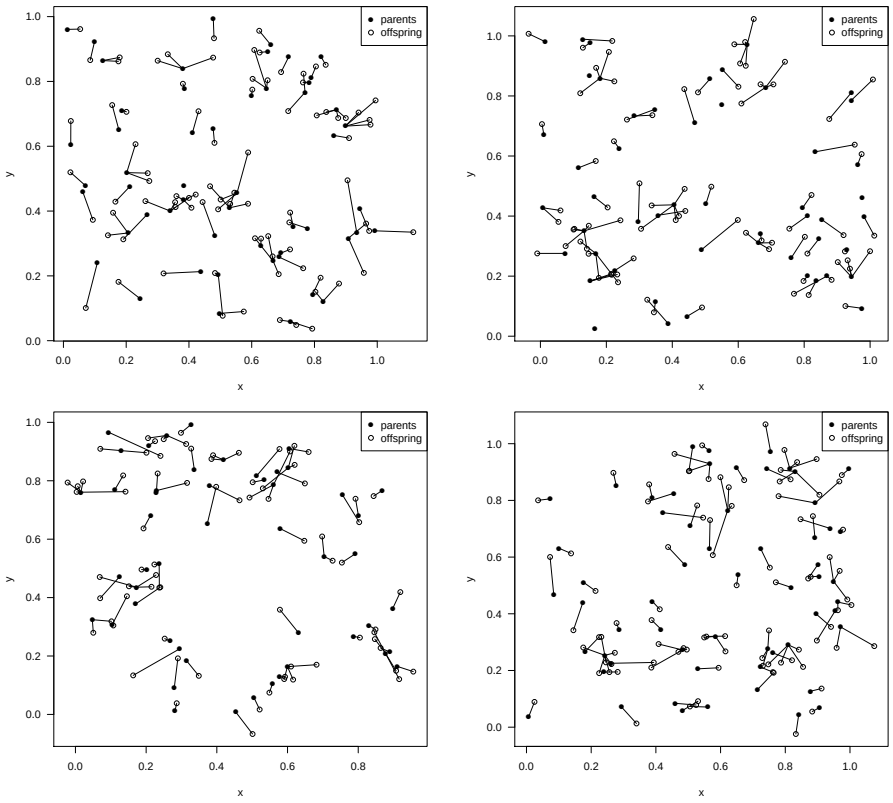


Figure 3. Sample simulations.

We first simulate the parents according to a homogeneous Poisson process with intensity $\gamma$ being the number of base points in the data set. The number of offspring per parent is generated from a one-inflated Poisson distribution with the probability of being in the one-component equal to the proportion of parents with one offspring; the Poisson component intensity is set to the mean number of offspring from parents with more than one offspring. The squared distrance betwen parents and their corresponding offspring are simulated according to an exponential distribution with rate $\theta_1$ where $\theta_1$ is obtained using Bayesian estimation with a gamma prior. The angle between the parent and offspring is simulated according to a von Mises mixture with parameters estimated using the `mix.vmf` function in the `Directional` `R` package. Figure 3 displays sample simulations for the subject displayed in Figure 2.

# 4 Estimating the parameters of the intensity function

Assuming the primary point locations $(\mathbf{c})$, the secondary point locations $(\mathbf{x})$ and the labels $\mathbf{m}$ are observed, the parametric inference under the model consists of estimating the

expected number of offsprings $\gamma$ and the kernel parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)$.

### 4.1 Kernel parameters

We use a conjugate gamma prior for the distance parameter $\theta_1$; Using the conjugate prior for each of the von Mises distributions (Guttorp and Lockhart, 1988) and a beta prior for the mixing probability. It is straightforward (Appendix C) to calculate a Gibbs sampler for the kernel parametrs $\boldsymbol{\theta}_{-1}$.

### 4.2 Expected cluster size

Next, we update the expected number of offsprings. Assume the prior on $\gamma$ is $p(\gamma) \propto \gamma^{\alpha-1}\exp(-\beta\gamma)$ we calculate

$$\begin{aligned} p(\gamma, (x), (c), \alpha, \beta, \boldsymbol{\theta}) &\propto p((x), \mathbf{m} \mid (c), \boldsymbol{\gamma}, \boldsymbol{\theta}) p(\gamma \mid \alpha, \beta) \\ &= \gamma^{\alpha-1+\sum 1(m_i=l)} \exp(-\gamma(\beta + \int_B k(\xi, c_l \mid \boldsymbol{\theta}) d\xi)), \end{aligned} \tag{8}$$

i.e., a gamma distribution with parameters $\alpha + \sum 1(m_i = l)$ and $\beta + \int_B k(\xi, c_l \mid \boldsymbol{\theta})d\xi$.

## 5 Estimating the properties of the cluster process

Being able to estimate the parameters of the model, we can now continue towards determining the labels (which secondary points belong to each of the primary points), and the more difficult problem of identifying from a set of points which are the primary ones.

### 5.1 What points belong to what cluster?

Knowing what points are primary and secondary, and having parametric models for the distance distribution and the cluster size distribution, how do we go about determining the most likely configuration of clusters? Note that this is just a preliminary step towards the squinting problem, since we assume that we know the cluster centers.

It follows from (7) that the marginal distribution of labels (connecting secondary points to primary points) is

$$\mathbf{P}_m(m = l \mid \psi, \boldsymbol{\theta}) = \frac{\int_B k(\xi, c_l \mid \boldsymbol{\theta})d\xi}{\sum_{n=1}^{n(\psi)} \int_B k(\xi, c_n, \boldsymbol{\theta})d\xi}. \tag{9}$$

In order to update the labels we use the full conditional distribution (Baddeley, 2010, Ex. 213):

$$\mathbf{P}(m_i = l \mid x_i, \mathbf{c}, \gamma, \boldsymbol{\theta}) \propto \gamma k(x_i, c_l \mid \boldsymbol{\theta})$$

or more precisely

$$\mathbf{P}(m_i = l \mid x_i, \mathbf{c}, \gamma, \boldsymbol{\theta}) = \frac{k(x_i, c_l \mid \theta_l)}{\sum_{n=1}^{n(\psi)} k(x_i, c_n \mid \boldsymbol{\theta})} \tag{10}$$

Figure 4 shows the high posterior density estimates of clusters. 62% of the labels agree with the data.

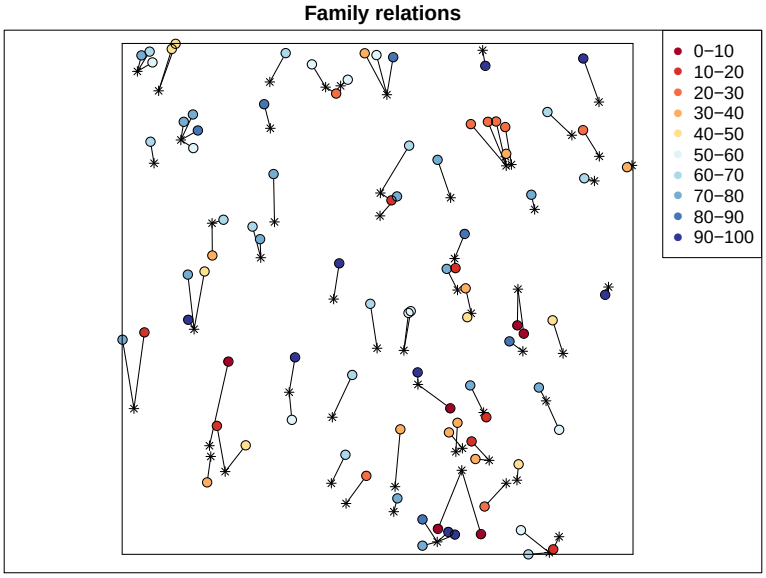


Figure 4. The posterior probability of each primary point being correctly assigned to its cluster.

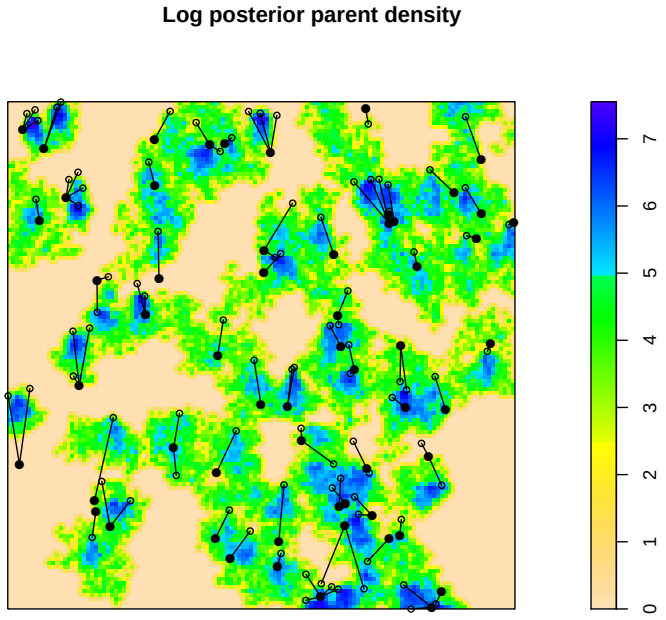


Figure 5. The log posterior parent density is depicted using the colour scale. The observed clusters are shown in black, where the secondary points are open circles and the primary are filled circles.

### 5.2 Identifying the cluster centers

If we want to update the cluster center locations we set down a prior for $\mathbf{c}$ proportional to $\lambda^{n(\psi)}$ and write

$$p(\psi) = p(\boldsymbol{\gamma}, \mathbf{c}) = p(\mathbf{c})p(\boldsymbol{\gamma} \mid \mathbf{c}) \propto \lambda^{n(\psi)} p(\boldsymbol{\gamma} \mid \mathbf{c})$$

to get

$$\begin{aligned} p(\mathbf{c} \mid \mathbf{x}, \mathbf{m}, \gamma, \boldsymbol{\theta}, \lambda) &\propto p(\mathbf{x}, \mathbf{m} \mid \mathbf{c}, \gamma, \boldsymbol{\theta}) p(\mathbf{c}) \\ &\propto \exp(n(\psi)|B| - \gamma \sum_{m=1}^{n(\psi)} \int_B k(\xi, c_l \mid \theta_l) d\xi) \Big( \prod_{(x_i, m_i) \in (\mathbf{x}, \mathbf{m})} \gamma k(x_i, c_{m_i} \mid \boldsymbol{\theta}) \Big) \lambda^{n(\psi)} \end{aligned} \quad (11)$$

An estimate of the parent density is shown in Figure 5. In most cases the posterior density is high at the actual cluster point. There are two instances where the algorithm did not find the actual location highly likely.

## 6 Squinting

The squinting approach to clustering looks for points that are close together. Following the likelihood-based approach we first determine the offspring points that are closest to the given cluster centers. The result is shown in Figure 6. 62% of the connections agree with the data.

In order to determine the best squinting clusters we use the K-means clustering algorithm (Hartigan and Wong, 1979). We let $K = 58$, using a random starting point for the algorithm. Figure 7 shows the observed fit in black and the k-means fit in red. There are 34 observed size 1 clusters, and the K-means algorithm yields 35. There are 18 observed size 2 clusters, and K-means yield 16. The largest observed clusters are size 4, while the largest K-means cluster is size 5.

## 7 Discussion

The main reason for using a Poisson cluster process with Poisson cluster size is that the likelihood is tractable. There are instances in other data sets where a Poisson process in not a particularly good model for the primary points. We can write out the likelihood for some other models, such as Matérn type III (see Guttorp and Thorarinsdottir (2011) for this and other examples). It would be interesting to investigate other tools, such as Palm likelihood methods or ABC, but it is beyond the scope of this paper.

In any case where we can write down a likelihood we could use it for estimating parameters, and apply EM techniques to estimate the unobserved links and primary points. The estimation and illustration of uncertainty for such non-Bayesian approaches is difficult.

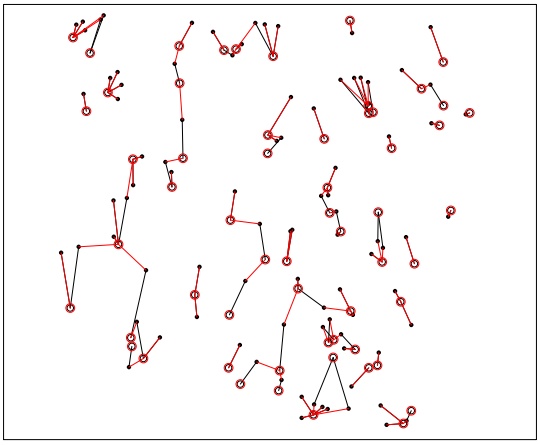

Figure 6. The offspring points closest to the cluster centers are connected to the cluster centers with red lines. The observed cluster structure is shown in black.

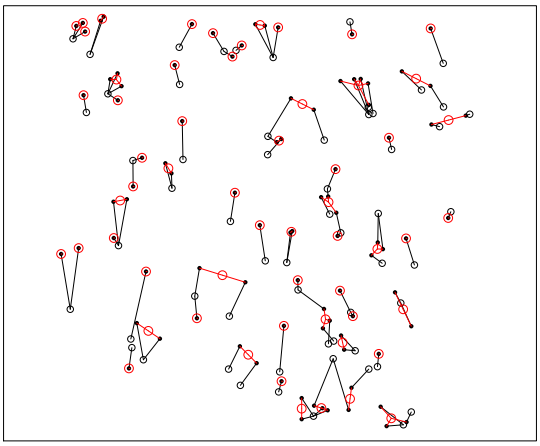

Figure 7. The K-means clustering (red) compared to the observed cluster structure (black).

We have illustrated Bayesian techniques for estimating the cluster structure of a cluster process, and applied the techniques to a particular data set where we actually know the true structure. While the reconstruction is far from perfect, it is remarkable that one can get as close as we do to the actual truth. The squinting approach is unable to deal with directional aspects of the process, and in that sense we find that Bayes does, indeed, beat squinting.

## Acknowledgments

We acknowledge the support of The Research Council of Norway through grant 240838 "Model selection and model verification for point processes".

## References

Albert-Green, A. (2016). *Joint Models for Spatial and Spatio-Temporal Point Processes*. PhD thesis, University of Western Ontario. 5

Andersson, C., Guttorp, P., and Särkkä, A. (2016). Discovering early diabetic neuropathy from epidermal nerve fiber patterns. *Statistics in Medicine*, 35:4427–4442. 4

Baddeley, A. (2010). Multivariate and marked point processes. In Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P., editors, *Handbook of Spatial Statistics*, chapter 21, pages 371–402. CRC Press, Boca Raton. 5, 6, 8

Guttorp, P. (1995). *Stochastic modeling of scientific data*. Capman & Hall, London. 4

Guttorp, P. and Lockhart, R. A. (1988). On finding the location of a signal: a Bayesian analysis. *Journal of the American Statistical Association*, 83:322–330. 8

Guttorp, P. and Thorarinsdottir, T. (2011). Bayesian inference for non-Markovian point processes. In Porcu, E., Montero, J., and Schlather, M., editors, *Space-Time Processes and Challenges Related to Environmental Problems: Proceedings of the Spring School "Advances and Challenges in Space-Time Modelling of Natural Events"*, pages 79–102. Springer, Berlin. 10

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28:100–108. 10

Kennedy, W., Wendelschafer-Crabb, G., and Johnson, T. (1996). Quantitation of epidermal nerves in diabetic neuropathy. *Neurology*, 47:1042–1048. 5

Møller, J. and Ghorbani, M. (2012). Aspects of Second-Order Analysis of Structured Inhomogeneous Spatio-Temporal Point Processes. *Statistica Neerlandica*, 66(4):472–491. 5

Neyman, J., Scott, E. L., and Shane, C. D. (1956). Statistics of images of galaxies with particular reference to clustering. In Neyman, J., editor, *Proc. Third Berkeley Symposium on Probability and Statistics, Volume III, Contributions to astronomy and physics*, pages 75–111. University of California Press, Berkeley and Los Angeles. 4

Prokešová, M. and Dvořák, J. (2014). Statistics for Inhomogeneous Space-Time Shot-Noise Cox Processes. *Methodology and Computing in Applied Probability*, 16(2):433–449. 5

Waller, L. A., Särkkä, A., Olsbo, V., Myllymäki, M., Panoutsopoulou, I. G., Kennedy, W. R., and Wendelschafer-Crabb, G. (2011). Second-order spatial analysis of epidermal nerve fibers. *Statistics in Medicine*, 30:2827–2841. 4

Wiegand, T., Gunatilleke, S., Gunatilleke, N., and Okuda, T. (2007). Analyzing the Spatial Structure of a Sri Lankan Tree Species with Multiple Scales of Clustering. *Ecology*, 88(12):3088–3102. 5

# A Calculating the density normalizing constant

We consider here the dispersion density for a parent point $c$ and offspring $x$. We assume that the squared distance between $x$ and $c$ follows an exponential distribution with parameter $\tau$ and that the offsprings spread around $c$ at a preferred angle according to a von Mises distribution with mean angle $\theta$ and spread parameter $\kappa$. For notational convenience, assume that $c = (0, 0)$. The dispersion density is then given by

$$\begin{aligned} k(x|c, \theta, \kappa, \tau) &= k(r\cos\nu, r\sin\nu|c, \theta, \kappa, \tau) \\ &= \frac{\exp(\kappa\cos(\theta - \nu))}{\pi I_0(\kappa)}\tau\exp(-\tau r^2). \end{aligned} \tag{A.1}$$

This is a kernel,

$$\begin{aligned} \int_2 k(x|c, \theta, \kappa, \tau)\mathrm{d}x &= \int_0^{2\pi}\int_0^{\infty} k(r\cos\nu, r\sin\nu|c, \theta, \kappa, \tau)r\mathrm{d}r\mathrm{d}\nu \\ &= \int_0^{2\pi}\frac{\exp(\kappa\cos(\theta - \nu))}{2\pi I_0(\kappa)}\int_0^{\infty}\tau\exp(-\tau r^2)2r\mathrm{d}r\mathrm{d}\nu \\ &= 1. \end{aligned}$$

To obtain the normalizing constant of the likelihood, we are interested in integrating the dispersion density over the observation window $B = [0, 1] \times [0, 1]$ for $x, c \in B$. To simplify the notation in our calculations, we shift the observation window by $c = (c_1, c_2)$ and integrate over $B_c = [-c_1, 1 - c_1] \times [-c_2, 1 - c_2]$ instead. To integrate over $B_c$ in polar coordinates, we need to integrate over four separate sets that combined constitute $B_c$,

$$\begin{aligned} \nu \in \Big(0, \frac{\pi}{4}\Big) \cup \Big(\frac{7\pi}{4}, 2\pi\Big), &\qquad r \in \big(0, (1 - c_1)\sec\nu\big) \\ \nu \in \Big(\frac{\pi}{4}, \frac{3\pi}{4}\Big), &\qquad r \in \big(0, (1 - c_2)\csc\nu\big) \\ \nu \in \Big(\frac{3\pi}{4}, \frac{5\pi}{4}\Big), &\qquad r \in \big(0, -c_1\sec\nu\big) \\ \nu \in \Big(\frac{5\pi}{4}, \frac{7\pi}{4}\Big), &\qquad r \in \big(0, -c_2\csc\nu\big) \end{aligned}$$

Straight forward calculations then give

$$\begin{aligned} \int_{B_c} k(x|(0, 0), \theta, \kappa, \tau)\mathrm{d}x = 1 - \frac{1}{2\pi I_0(\kappa)}\Big[ &\int_{-\pi/4}^{\pi/4}\exp\big(\kappa\cos(\theta - \nu) - \tau(1 - c_1)^2\sec^2\nu\big)\mathrm{d}\nu \\ &+ \int_{\pi/4}^{3\pi/4}\exp\big(\kappa\cos(\theta - \nu) - \tau(1 - c_2)^2\csc^2\nu\big)\mathrm{d}\nu \\ &+ \int_{3\pi/4}^{5\pi/4}\exp\big(\kappa\cos(\theta - \nu) - \tau c_1^2\sec^2\nu\big)\mathrm{d}\nu \\ &+ \int_{5\pi/4}^{7\pi/4}\exp\big(\kappa\cos(\theta - \nu) - \tau c_2^2\csc^2\nu\big)\mathrm{d}\nu\Big]. \end{aligned}$$

# B Inference for kernel parameters with known parents

Assume the observation window is $B = [0,1]^2$, let $M$ denote the number of i.i.d. clusters and $N$ the number of data points. Define

$$A(\boldsymbol{\theta}) := \int_B \gamma \sum_{m=1}^{M} k(\xi, c_m \,|\, \boldsymbol{\theta}) \mathrm{d}\xi$$

and

$$G(u, \boldsymbol{\theta}) := \Big[\theta_2 \frac{\exp(\theta_4 \cos(u - \theta_3))}{2\pi I_0(\theta_4)} + (1-\theta_2)\frac{\exp(\theta_6 \cos(u-\theta_5))}{2\pi I_0(\theta_6)}\Big].$$

The likelihood is then given by

$$\begin{aligned} p((\mathbf{x}, \mathbf{m}) \,|\, \psi, \boldsymbol{\theta}) &= \exp(M - A(\theta)) \prod_{i=1}^{N} \gamma k(x_i, c_{m_i} \,|\, \boldsymbol{\theta}) \\ &= \exp(M - A(\boldsymbol{\theta})) \prod_{i=1}^{N} \Big[\gamma\theta_1 \exp(-\theta_1 r_i) G(u_i, \boldsymbol{\theta})\Big], \end{aligned}$$

where $r_i = \|x_i - c_{m_i}\|^2$ and $u_i = \angle(x_i, c_{m_i})$. It follows that

$$\log p((\mathbf{x}, \mathbf{m}) \,|\, \psi, \boldsymbol{\theta}) = M - A(\boldsymbol{\theta}) + N \log\gamma + N \log\theta_1 - \sum_{i=1}^{N} \theta_1 r_i + \sum_{i=1}^{N} \log G(u_i, \boldsymbol{\theta}).$$

Due to the complicated form of the likelihood, all components of $\boldsymbol{\theta}$ must be estimated with a Metropolis Hastings algorithm. Denote the current parameter value by $\boldsymbol{\theta}$. For $j = 1, \ldots, 6$, let $q_j(\cdot \,|\, \boldsymbol{\theta})$ denote the proposal distribution for the $j$-th component of $\boldsymbol{\theta}$. We sample a new value $\theta_j^*$ from $q_j$ and denote by $\boldsymbol{\theta}^{(j)*}$ the vector with $\theta_j$ replaced by $\theta_j^*$. We then accept the new proposal $\theta_j^*$ with probability

$$\min\Big\{\frac{p((\mathbf{x}, \mathbf{m}) \,|\, \psi, \boldsymbol{\theta}^{(j)*}) q_j(\theta_j \,|\, \boldsymbol{\theta}^{(j)*})}{p((\mathbf{x}, \mathbf{m}) \,|\, \psi, \boldsymbol{\theta}) q_j(\theta_j^* \,|\, \boldsymbol{\theta})}, 1\Big\}.$$

Usually, we calculate the ratio on a log scale,

$$\log p((\mathbf{x}, \mathbf{m}) \,|\, \psi, \boldsymbol{\theta}^{(j)*}) - \log p((\mathbf{x}, \mathbf{m}) \,|\, \psi, \boldsymbol{\theta}) + \log q_j(\theta_j \,|\, \boldsymbol{\theta}^{(j)*}) - \log q_j(\theta_j^* \,|\, \boldsymbol{\theta}).$$

Note that for $j = 1$,

$$\begin{aligned} &\log p((\mathbf{x}, \mathbf{m}) \,|\, \psi, \boldsymbol{\theta}^{(j)*}) - \log p((\mathbf{x}, \mathbf{m}) \,|\, \psi, \boldsymbol{\theta}) \\ &\qquad = -A(\boldsymbol{\theta}^{(j)*}) + A(\boldsymbol{\theta}) + N(\log\theta_1^* - \log\theta_1) - \sum_{i=1}^{N} (\theta_1^* - \theta_1) r_i, \end{aligned}$$

while for $j > 1$,

$$\begin{aligned} &\log p((\mathbf{x}, \mathbf{m}) \,|\, \psi, \boldsymbol{\theta}^{(j)*}) - \log p((\mathbf{x}, \mathbf{m}) \,|\, \psi, \boldsymbol{\theta}) \\ &\qquad = -A(\boldsymbol{\theta}^{(j)*}) + A(\boldsymbol{\theta}) + \sum_{i=1}^{N} \log G(u_i, \boldsymbol{\theta}^{(j)*}) - \log G(u_i, \boldsymbol{\theta}). \end{aligned}$$

# C Allowing a mixture of von Mises distributions

### Rewrite mixture density in canonical exponential family form:

$$f(y) = \sum_{j=1}^{2} p_j \frac{\exp\left\{\tau_j^T y\right\}}{2\pi \mathrm{I_0}(|\tau_j|)}$$

where: $|\tau_j| = \left(\tau_j^T \tau_j\right)^{1/2}$, $\tau_j^T = \kappa_j(\cos\mu_j,\ \sin\mu_j)$, $Y = (\cos\Theta,\ \sin\Theta)$.

### Prior distributions:

- $f(p_j) = \text{Beta}(\delta_1, \delta_2)$
- $f(\tau_j) \propto \frac{\exp\{\tau_j^T y_o\}}{\mathrm{I_0}(|\tau_j|)^c}$, where:
  - $y_0 = R_0(\cos\theta_0,\ \sin\theta_0)$ is a prior observation
  - $c$ is the prior sample size
  - $R_0$ is the prior parameter representing the component on the $x$-axis (in the known direction) of the $c$ observations
  - $\theta_0$ is the prior observation expressed as an angle
- The normalizing constant of $f(\tau)$ is $H_1(R_0, c)^{-1}$ where:
  - $H_\alpha(\beta, \gamma) = \int_0^\infty x^\alpha \mathrm{I_0}(\beta x)\mathrm{I_0}(x)^{-\gamma}\mathrm{d}x$, so $H_1(R_0, c) = \int_o^\infty x\mathrm{I_0}(R_0 x)\mathrm{I_0}(x)^{-c}\mathrm{d}x$ and $H_1(R_0, c)^{-1} = \left[\int_o^\infty x\mathrm{I_0}(R_0 x)\mathrm{I_0}(x)^{-c}\mathrm{d}x\right]^{-1}$
- Therefore,

$$f(\tau) = \frac{\exp\left\{\tau^T y_0\right\}}{\mathrm{I_0}(|\tau_j|)^c \int_0^\infty x\mathrm{I}_0(R_0 x)\mathrm{I_0}(x)^{-c}\mathrm{d}x}$$

### Indicators for missing data on component membership:

Introduce indicators $\boldsymbol{z} = \{z_{ij},\ i = 1, 2, \ldots, n,\ j = 1, 2\}$ where

$$z_{ij} = \begin{cases} 1, & \text{if } i\text{th observation, } Y_i, \text{ belongs to } j\text{th component} \\ 0, & \text{otherwise} \end{cases}$$

- $f(y_i \mid z_{ij} = 1) = \text{VM}(\mu_j,\ \tau_j),\ f(z_{ij} = 1 \mid p) = p_j$
- $z_{ij} \sim \text{Bernoulli}(p_j), \forall \boldsymbol{z}_i = \{z_{ij}\}$
- $\boldsymbol{z}_i \mid \boldsymbol{p} \sim \text{Multinomial}(1, p_1, p_2)$

### Joint distribution of observed data ($y$) and unobserved data ($z$)

$$f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\tau}, \boldsymbol{p}) \quad = \quad f(\boldsymbol{y} \mid \boldsymbol{\tau}, \boldsymbol{z}) f(\boldsymbol{z} \mid \boldsymbol{p})$$

where:

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{\tau}, \boldsymbol{z}) &= \prod_{i=1}^{n}\prod_{j=1}^{2}\left(\frac{\exp\{\tau_j^T y\}}{2\pi \mathrm{I_0}(|\tau_j|)}\right)^{z_{ij}} \\ f(\boldsymbol{z} \mid \boldsymbol{p}) &= \prod_{i=1}^{n}\prod_{j=1}^{2}(p_j)^{z_{ij}} \end{aligned}$$

so:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\tau}, \boldsymbol{p}) &= \prod_{i=1}^{n}\prod_{j=1}^{n}\left(\frac{\exp\{\tau_j^T y\}}{2\pi \mathrm{I_0}(|\tau_j|)}\right)^{z_{ij}} \prod_{i=1}^{n}\prod_{j-1}^{2}(p_j)^{z_{ij}} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{n}\left[p_j \frac{\exp\{\tau_j^T y\}}{2\pi \mathrm{I_0}(|\tau_j|)}\right]^{z_{ij}} \end{aligned}$$

## Joint distribution of the parameters:

$$\begin{aligned} f(\boldsymbol{\tau}, \boldsymbol{p} \mid \boldsymbol{y}, \boldsymbol{z}) &\propto f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\tau}, \boldsymbol{p}) \\ &= f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{\tau}) f(\boldsymbol{z} \mid \boldsymbol{p}) f(\boldsymbol{\tau}) f(\boldsymbol{p}) \end{aligned}$$

where:

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{\tau}) f(\boldsymbol{z} \mid \boldsymbol{p}) &= \prod_{i=1}^{n}\prod_{j=1}^{2}\left[p_j \frac{\exp\{\tau_j^T y\}}{2\pi \mathrm{I_0}(|\tau_j|)}\right]^{z_{ij}} \\ f(\boldsymbol{\tau}) &= \prod_{j=1}^{2} \frac{\exp\left\{\tau_j^T y_0\right\}}{\mathrm{I_0}(|\tau_j|)^c \int_0^\infty x \mathrm{I_0}(R_0 x)\mathrm{I_0}(x)^{-c}\mathrm{d}x} \\ f(\boldsymbol{p}) &= \frac{\Gamma(\sum_{j=1}^{2}\delta_j)}{\prod_{j=1}^{2}\Gamma(\delta_j)} \prod_{j=1}^{2} p_j^{\delta_j - 1} \end{aligned}$$

Putting this all together, the joint distribution is:

$$\begin{aligned} f(\boldsymbol{\tau}, \boldsymbol{p} \mid \boldsymbol{y}, \boldsymbol{z}) &\propto f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{\tau}) f(\boldsymbol{z} \mid \boldsymbol{p}) f(\boldsymbol{\tau}) f(\boldsymbol{p}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{2}\left[p_j \frac{\exp\{\tau_j^T y\}}{2\pi \mathrm{I_0}(|\tau_j|)}\right]^{z_{ij}} \frac{\exp\left\{\tau_j^T y_0\right\}}{\mathrm{I_0}(|\tau_j|)^c \int_0^\infty x \mathrm{I_0}(R_0 x)\mathrm{I_0}(x)^{-c}\mathrm{d}x} \frac{\Gamma(\sum_{j=1}^{2}\delta_j)}{\prod_{j=1}^{2}\Gamma(\delta_j)} \prod_{j=1}^{2} p_j^{\delta_j - 1} \end{aligned}$$

## Full Conditional Posterior for Gibbs Sampler:

**von Mises Canonical parameter, $\tau_j$:**

$$\begin{aligned} f(\tau_1 \mid \tau_2, \boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z}) &\propto f(\tau_2, \boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z}) f(\tau_1) \\ &= f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{\tau}) f(\boldsymbol{z} \mid \boldsymbol{p}) f(\tau_1) f(\tau_2) f(\boldsymbol{p}) \\ &\propto f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{\tau}) f(\boldsymbol{\tau}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{2}\left[\frac{\exp\{\tau_j^T y\}}{2\pi \mathrm{I_0}(|\tau_j|)}\right]^{z_{ij}} \frac{\exp\left\{\tau_j^T y_{0i}\right\}}{\mathrm{I_0}(|\tau_j|)^c \int_0^\infty x \mathrm{I_0}(R_0 x)\mathrm{I_0}(x)^{-c}\mathrm{d}x} \\ &\propto \prod_{i=1}^{n}\left[\frac{\exp\{\tau_1^T y\}}{2\pi \mathrm{I_0}(|\tau_1|)}\right]^{z_{i1}} \frac{\exp\left\{\tau_1^T y_{0i}\right\}}{\mathrm{I_0}(|\tau_1|)^c \int_0^\infty x \mathrm{I_0}(R_0 x)\mathrm{I_0}(x)^{-c}\mathrm{d}x} \end{aligned}$$

and

$$\begin{aligned} f(\tau_2 \mid \tau_2, \boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z}) &\propto f(\tau_1, \boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z}) f(\tau_2) \\ &= \prod_{i=1}^{n} \left[ \frac{\exp\{\tau_2^T y\}}{2\pi \mathrm{I_0}(|\tau_2|)} \right]^{z_{i2}} \frac{\exp\left\{\tau_2^T y_{0i}\right\}}{\mathrm{I_0}(|\tau_2|)^c \int_0^\infty x \mathrm{I_0}(R_0 x) \mathrm{I_0}(x)^{-c} \mathrm{d}x} \end{aligned}$$

**Mixture probability, $p$:**

$$\begin{aligned} f(\boldsymbol{p} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\tau}) &\propto f(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\tau} \mid \boldsymbol{p}) f(\boldsymbol{p}) \\ &= f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{\tau}) f(\boldsymbol{z} \mid \boldsymbol{p}) f(\boldsymbol{\tau}) f(\boldsymbol{p}) \\ &\propto f(\boldsymbol{z} \mid \boldsymbol{p}) f(\boldsymbol{p}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{2} p_j^{z_{ij}} \frac{\Gamma(\sum_{j=1}^{2} \delta_j)}{\prod_{j=1}^{2} \Gamma(\delta_j)} \prod_{j=1}^{2} p_j^{\delta_j - 1} \\ &= \prod_{j=1}^{2} p_j^{\sum_i z_{ij} + \delta_j - 1} \frac{\Gamma(\sum_{j=1}^{2} \delta_j)}{\prod_{j=1}^{2} \Gamma(\delta_j)} \end{aligned}$$

If $\delta_1 = \delta_2 = 1$:

$$\begin{aligned} \prod_{j=1}^{2} p_j^{\sum_i z_{ij} + \delta_j - 1} \frac{\Gamma(\delta_1 + \delta_2)}{\Gamma(\delta_1)\Gamma(\delta_2)} &= \prod_{j=1}^{2} p_j^{\sum_i z_{ij} + \delta_j - 1} \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} \\ &= \prod_{j=1}^{2} p_j^{\left(\sum_i z_{ij} + 1\right) - 1} \\ &= \text{Dirichlet}(1 + \textstyle\sum_i z_{i1},\ 1 + \sum_i z_{i2}) \\ &= \text{Beta}(1 + \textstyle\sum_i z_{i1},\ 1 + \sum_i z_{i2}) \end{aligned}$$

**Missing data indicator, $\boldsymbol{z}_i = (z_{i1}, z_{i2})$:**

$$\begin{aligned} f(z_{ij} = 1 \mid y_i, \boldsymbol{\tau}, \boldsymbol{p}) &= \frac{f(y_i \mid \boldsymbol{\tau}, \boldsymbol{p}, z_{ij} = 1) f(z_i = 1 \mid \boldsymbol{\tau}, \boldsymbol{p})}{\sum_{j=1}^{2} f(y_i \mid \boldsymbol{\tau}, p, z_{ij} = 1) f(z_i = 1 \mid \boldsymbol{\tau}, \boldsymbol{p})} \\ &= \frac{f(y_i \mid \boldsymbol{\tau}, z_{ij} = 1) f(z_i = 1 \mid \boldsymbol{p})}{\sum_{j=1}^{2} f(y_i \mid \boldsymbol{\tau}, z_{ij} = 1) f(z_i = 1 \mid \boldsymbol{p})} \\ &= \frac{f(y_i \mid \boldsymbol{\tau}, z_{ij} = 1) p_j}{\sum_{j=1}^{2} f(y_i \mid \boldsymbol{\tau}, z_{ij} = 1) p_j} \\ &= \frac{f(y_i \mid \boldsymbol{\tau}, z_{ij} = 1) p_j}{f(y_i)} \end{aligned}$$

$$\begin{aligned} f(z_{ij} = 1 \mid y_i, \boldsymbol{\tau}, \boldsymbol{p}) &= 1 - f(z_{ij} = 0 \mid y_i, \boldsymbol{\tau}, \mathbf{p}) \\ f(z_{ij} = 0 \mid y_i, \boldsymbol{\tau}, \mathbf{p}) &= 1 - \frac{f(y_i \mid \boldsymbol{\tau}, z_{ij} = 1) p_j}{f(y_i)} \end{aligned}$$

Therefore, $Z_{ij} \sim \text{Bernoulli}\left(\frac{f(y_i|\boldsymbol{\tau}, z_{ij}=1)p_j}{f(y_i)}\right)$ and $\boldsymbol{Z}_i = \{Z_{ij}\} \sim \text{Multinomial}(1, w_{i1}, w_{i2})$ where $w_{ij} = \frac{f(y_i|\tau_j, z_{ij}=1)p_j}{f(y_i)}$.