Proceedings of the Federated Conference on Computer Science and Information Systems pp. 481–484
DOI: 10.15439/2017F264
ISSN 2300-5963 ACSIS, Vol. 11

Catching clouds: Simultaneous optimization of the parameters of biological agent plumes using Dirichlet processes to best estimate infection source location

James Thompson, Thomas Finnie, Ian Hall
Public Health England
Porton Down
Salisbury, SP4 0JG
United Kingdom
Email: thomas.finnie@phe.gov.uk

Nina Dobrinkova
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
acad. Georgi Bonchev bl. 2
Bulgaria
Email: ninabox2002@gmail.com

This work has been funded by: the EC Research Executive Agency 7th Framework Programme, (SEC-2013.4.1-4) under grant number FP7-SEC-2013-608078-IMproving Preparedness and Response of HEalth Services in major criseS (IMPRESS), the UK NIH Health Protection Research Unit in Emergency Preparedness and Response and by grant 02/20 awarded from the Bulgarian National Science Fund.

***Abstract*—We describe a stochastic method using Dirichlet processes to derive mixture models that allow the numerical description of outbreaks of diseases with multiple sources. We show that existing disease models may be extended using this method and how this may be used in a practical context to support the simulated response to a mass casualty public health emergency.**

***Index terms* epidemiology, stochastic processes, clustering, mixture models, Chinese restaurant process, non-parametric fitting.**

I. Introduction

MODERN epidemiological practice during the investigation of the outbreak of a disease often involves the construction of mathematical or computation models. Frequently, these models are used to answer operational questions such as forecasting where further cases are likely to occur, what the total number of casualties are likely to be, and perhaps even where the likely source of the disease may be found. To be of greatest value such estimates need to be made from a small number of cases early in the course of the outbreak so that public health officials may prioritize resources.

Many of these epidemiological models use the approach of representing the outbreak as a *probability density function* [1]–[3]. As such, samples taken from this function should produce a similar spread of cases as the real outbreak. These probability density functions are usually parametrized by a fixed and finite number of parameters *e.g.* spatio-temporal location, climatic conditions, transmission rates, *etc*. [3]. The values of these parameters are manipulated until a set is found which maximises the likelihood of the probability density function. This optimisation process is a well known problem in mathematics and computer science with a huge literature. Major reviews may be found in [4]–[6] and also see Fletcher [7] for a partial overview and introduction of current theory and techniques.

Diseases however may not be limited to a single source or event. Examples where there may be multiple clusters within an outbreak could include:

- A legionella outbreak where multiple cooling towers or air conditioning units are responsible for the cases.
- A shipment of infected food distributed to a large number restaurants, schools, canteens *etc.* over a region.
- A terrorist incident involving multiple covert releases of a pathogen in a short period of time.

Where there are multiple sources, the scenario can be thought of as several simultaneous, independent outbreaks in space and time. Since we do not know *a priori* how many sources there are, the process of determining the parameter values becomes substantially more difficult. The problem now requires a solution related to cluster analysis for which there are many well-known algorithms such as *k-means clustering* [8], [9], *principal component analysis* [10]–[12] and *hierarchical cluster analysis* [13] which have been applied to problems within the field of epidemiology [14]–[17].

Applying such clustering algorithms directly to a multi-outbreak situation is complicated, especially as it may not be clear many sources of exposure there are and consequently what the 'correct' number of clusters should be. In this paper we show how multiple solutions for the value of parameters in the base model can be considered as a *mixture model*, *i.e.* a weighted sum of several probability density functions each with different parameter values. We show how such a mixture model may be calculated by applying a *Dirichlet process* and extend a single source disease plume model using Dirichlet processes to encompass multiple sources to provide a concrete example of this method in action during a table top exercise.

II. Mixture distributions

A. Finite mixture models

In this paper we denote probability density functions by F (or F_*) and probability mass functions by H (or H_*).

Given a finite collection of probability density functions, a *finite mixture distribution* is the probability density function for a random variable derived by first randomly selecting one of the probability density functions and then drawing a sample from that probability density function.

Formally, let $F_1, F_2, \ldots, F_n$ be probability density functions with the same domain and $w_1, w_2, \ldots, w_n$, be positive real numbers (weights) such that $w_1 + w_2 + \ldots + w_n = 1$.

Then the probability density function for the derived mixture distribution is given by:

$$F(x) = \sum_{i=1}^{n} w_i F_i(x).$$

To sample from this distribution we first choose a distribution F_k with probability $\mathbf{P}(k = i) = w_i$, then we draw from F_k.

From an epidemiology perspective, we can think of each pair (F_i, w_i) as distinct clusters and the probability that a case belongs to that cluster. Here we take all the probability density functions to be the same (*e.g.* all lognormal distributions), but this is not necessary.

B. Infinite mixture models

The mixture model can be extended in a natural way to an *infinite mixture distribution*. Infinite mixtures often have much nicer theoretical properties than finite mixtures and in the next section we describe a natural relationship between infinite mixtures and Dirichlet processes. In particular, infinite mixtures models are often used as they allow us to "*by-pass the need to determine the "correct" number of components in a finite mixture model, a task fraught with technical difficulties*" [18].

An infinite mixture distribution is defined to be:

$$F(x) = \sum_{i=1}^{\infty} w_i F_i(x),$$

note that we still require that

$$\sum_{i=1}^{\infty} w_i = 1.$$

As before, we sample from this distribution by first choosing a distribution F_k with probability $\mathbf{P}(k = i) = w_i$, then sampling from F_k. In practice it is computationally impossible to construct an infinite mixture model, instead we approximate them with finite mixtures for some very large n.

III. Dirichlet processes

A. A formal definition of a Dirichlet process

A *stochastic process* is a distribution over a function space. Each sample path from the stochastic process is a function drawn from the distribution. *Dirichlet processes* are a class of stochastic process where the sample path is a probability distribution with special properties. Less formally, a Dirichlet process is a distribution over distributions, and draws from a Dirichlet process are random probability mass functions.

Dirichlet processes can be thought of as an infinite dimensional generalization of the Dirichlet distribution. Recall (from [19]) that the Dirichlet distribution $\boldsymbol{Dir}(\boldsymbol{\alpha})$ is a continuous multivariate probability density function parametrized by K, the number of dimensions and a vector of K positive reals $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, the *concentration parameters*.

Let F (the *base distribution*) be a probability density function with support $\boldsymbol{S}$, and α (the *concentration parameter*) be a positive real number. We denote the Dirichlet process by $\boldsymbol{DP}(F, \alpha)$. F is the expected value of the Dirichlet process and draws from $\boldsymbol{DP}(F, \alpha)$ are 'around' F (in the same way that draws from a normal distribution are around the mean). It is impossible to describe $\boldsymbol{DP}(F, \alpha)$ itself or any probability mass function H drawn from $\boldsymbol{DP}(F, \alpha)$, both would require an infinite amount of information. However there are properties of $\boldsymbol{DP}(F, \alpha)$ and $H \sim \boldsymbol{DP}(F, \alpha)$ that can be precisely stated.

Let $\{S_i\}_{i=1}^{n}$ be a measurable finite partition of $\boldsymbol{S}$ and H be a random probability mass function distributed according to $\boldsymbol{DP}(F, \alpha)$ (remember that $\boldsymbol{DP}(F, \alpha)$ is a 'distribution of distributions'). Then the random vector

$$(H(A_1), \ldots, H(A_K)) \tag{1}$$

is distributed according to the multivariate distribution

$$\boldsymbol{Dir}(\alpha F(A_1), \ldots, \alpha F(A_K)). \tag{2}$$

Note that we have made no assumptions on the base probability density function F, in particular we have not assumed that it is parametrized, or even finitely parametrisable.

The concentration parameter α controls the 'discreteness' of the distributions drawn from $\boldsymbol{DP}(F, \alpha)$. As $\alpha \to 0$ the drawn distribution becomes more concentrated at a single value and at the limit the distribution is a Dirac delta function. As $\alpha \to \infty$ the drawn distribution becomes 'more continuous', and in the limit, the distributions are continuous *i.e.* they are probability density functions. Note that for finite α any distribution drawn from $\boldsymbol{DP}(F, \alpha)$ will *almost surely* be a probability mass function.

We cannot draw a distribution H explicitly from $\boldsymbol{DP}(F, \alpha)$. Instead we use a method that allows us to draw a large number of observations $X_1, X_2, \ldots$ from H without ever describing H concretely.

Given F and α as above, we sample $X_1, X_2, \ldots$ from H as follows:

1) Sample X_1 from F.
2) For $n > 1$:
   a) With probability $\frac{\alpha}{\alpha+n-1}$ draw X_n from F.
   b) With probability $\frac{n_i}{\alpha+n-1}$ set $X_n = X_i$, where n_i is the number of X_j, $j < n$ such that $X_j = X_i$.

It can be shown rigorously, using *de Finetti's theorem*, that this process is the same as drawing a probability mass function from $\boldsymbol{DP}(F, \alpha)$, then sampling $X_1, X_2, \ldots$ from H (see, for

example, Aldous [20]). This construction is often called the *Chinese restaurant process*.

B. From mixture models to Dirichlet processes

Our assumption was that complex multi-source disease outbreaks can be approximated by samples from *finite* mixture models, a weighted sum of finitely many parametrized distributions. Our goal is to find the parameters and the weights based on a small number of observations.

Example Consider a mixture model where the probability density function F is given by a sum of three 1-dimensional Gaussians with means and standard deviations μ_1, μ_2, μ_3 and $\sigma_1, \sigma_2, \sigma_3$ respectively. As μ_i are real numbers and σ_i are positive real numbers, the parameter space for F is $(\mathbb{R} \times \mathbb{R}_{>0})^3$ and the probability mass function is:

$$H(x) = \begin{cases} w_1, & \text{if } \mathrm{x} = (\mu_1, \sigma_1)\,; \\ w_2, & \text{if } \mathrm{x} = (\mu_2, \sigma_2)\,; \\ w_3, & \text{if } \mathrm{x} = (\mu_3, \sigma_3)\,; \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Given no information about F except for a small number of values sampled from it, we seek to recover H. However it is not possible to do this directly. In the next section we describe a method for approximating H using Dirichlet processes.

IV. Gibbs sampling in a Monte Carlo Markov chain

The approach used here is that of a modified Gibbs sampling algorithm on a Monte Carlo Markov chain (MCMC) and follows from algorithm 5 as described by Neal [18]. This method is specifically for use when we believe the unknown probability density function is a mixture model where all the components are from the same family of parameterized probability density functions.

Our hypothesis is that there exists a mixture distribution that explains the data. Since there is a bijection between mixture models and probability mass functions on the parameter space, we use Dirichlet processes to find the probability mass function corresponding to the 'best' mixture model.

Let $y_1, \ldots, y_n$ be our data and F_θ a distribution parameterized by θ and let G_0 be a base distribution on the parameter space Θ. Then our hypothesis can be restated as the data y_i are distributed identically to samples from the mixture distribution

$$F(y) = \sum_{i=1}^{\infty} w_i F_{\theta_i}(y).$$

The goal is to find the mixture, *i.e.* the pairs (w_i, θ_i), that best explain the data. This is equivalent to finding a probability mass function H over Θ.

The likelihood function is defined to be $\mathbf{F}(y_i, \theta) = F_\theta(y_i)$. We initialize the Markov chain by randomly sampling n times from G_0, *i.e.* we draw n sets of parameters $\{\theta_i\}_{i=1}^n$ from Θ parameter space.

We repeatedly sample from the MCMC as follows:

1) For each data point y_i, $i = 1, \ldots n$, update θ_i. First generate a candidate θ_i^* as follows:
   a) With probability $\frac{\alpha}{\alpha+n-1}$, use θ_i^* from G_0.
   b) With probability $\frac{n_i}{\alpha+n-1}$ set $\theta_i^* = \theta_j$, where $i \neq j$.

   The acceptance probability is

   $$\theta a(\theta_i^*, \theta_i) = \min\left\{1, \frac{\mathbf{F}(y_i, \theta_i^*)}{\mathbf{F}(y_i, \theta_i)}\right\}.$$

   With probability $\theta a(\theta_i^*, \theta_i)$, set θ_i to be equal to θ_i^*, otherwise leave θ_i unchanged. Repeat this step 'several times' (to ensure thorough mixing and thinning).
2) Update each distinct θ_i by drawing a new value from $\theta_i | y_i$.

After sampling from the MCMC chain a large number of times, normalize to get a probability mass function H over Θ. We then map this to a finite mixture model F. The probability mass function H, and consequently the mixture model, arising from this process is very likely to have a large number of components. This reflects the uncertainty arising from the small number of cases and the limitation of the models. The resulting probability mass function H is likely to have many thousands of components depending of the length of the chain. If the MCMC has converged we would expect the components of H to be grouped around the components of the 'true' mixture model where the 'true' mixture model is likely to consist of a small number of components, reflecting the small number of sources.

Since each state of the MCMC is an assignment of a set of parameters to each observed data point, we can take a 'vertical slice' through the MCMC chain. That is, we can isolate individual data points or subsets of data points and produce probability mass functions for data points of particular interest. This could also allow the additional weighting for specific data points. *e.g.* in an epidemiological context, we may be unsure about the diagnosis of some patients, while being sure about others. We could use this information to weight the more certain cases more heavily.

V. Example of use

This optimization method was developed to extend existing disease models to respond to the challenge of multiple source disease outbreaks. To ensure greatest utility of this method in genuine emergency situations the Dirichlet process optimizer was implemented as part of a wider suite of large-scale emergency response tools constructed as part of the IMPRESS system [21]. The IMPRESS system's components cover the range of emergency response disciplines from field triage through to strategic oversight of a broad scale biological incident. These capabilities are designed to strengthen coordination between response organizations and emergency medical services, including requests for international support.

Here the method was applied to an implementation of the Anthrax infection model presented in Legrand's 2009 paper [3]. The model describes a covert, aerosolized release of Anthrax in a populated area. The model parameters to be optimized are: the location of the release, the time of the release, the amount of Anthrax released, and parameters related to wind speed and wind direction at the time of release.

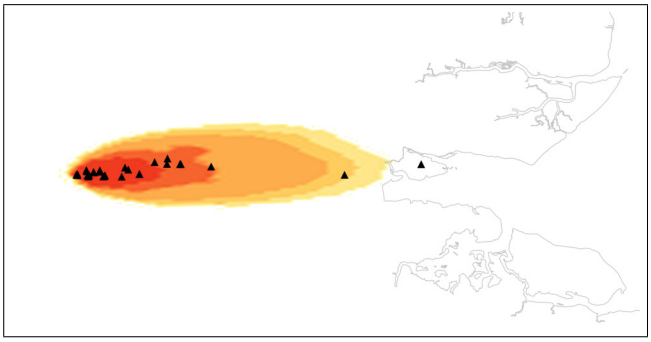

Fig. 1. A simulated example of the plume output from SorLoc. Input case locations are marked as triangles with output plume density shaded.

The optimizer combined with the disease model forms the SorLoc (source location) module.

The SorLoc module was tested and validated in the last phase of the IMPRESS project life cycle as part of a wider Greek-Bulgarian table top exercise. The Table Top Exercise was held in Sofia, Bulgaria on 16th March 2017. The exercise was operated by Greek and Bulgarian actors drawn from public services and hospitals across the two countries.

The exercise was based on a scenario where a combination of heavy rainfall and a strong earthquake had struck Southern Bulgaria. As a result, extensive damage was caused to buildings and infrastructure along with a landslide which damaged the road beside the Struma(BG)/Strimon(GR) River causing the river to overflow and flood part of the E79 Highway. These incidents were coupled with multiple car accidents caused by rockfalls along this segment of the E79 near the Greek-Bulgarian border. This situation generated many fatalities and injuries requiring immediate response, pre-hospital medical intervention and transportation of casualties to nearby hospitals. Victims' transportation via the collapsed E79 connecting the southern part of Bulgaria with the rest of the country caused the Bulgarian authorities to request international medical assistance, activating the standard procedures via the European Emergency Response Centre (EERC) in Brussels.

In order to facilitate the SorLoc module demonstration within this exercise, a scenario for aerosol released Anthrax was run in parallel to the main exercise. An outbreak was simulated for Shoreditch, London. We presented the course of the epidemic (as home locations and time at which each person fell ill) to SorLoc at a simulated five days from the first case and ran the optimization so that predictions of further evolution of the disease, numbers and locations of affected people and the original source of the outbreak might be calculated.

An illustrative input and result from the SorLoc module may be found in Figure 1. The output is provided as the inhalational dose generated by the plume(s) on a raster grid. This allows direct and immediate interpretation of size and the scale for the outbreak. It also provides a foundation for the mitigation effort and delivery of countermeasures to the population.

VI. Conclusion

Within this paper we have shown that a model which has been formulated as a probability density function and which would ordinarily be solved by standard optimization techniques may be extended to support multiple versions of the modeled process through the use of mixture models and Dirichlet processes. In addition we have explicitly shown the use of this within the field of disease modeling where it is directly applicable to existing models. We also constructed and demonstrated a production-ready implementation which was used to support a simulated response to a mass casualty, public health emergency.

References

[1] R. L. Prentice and R. Pyke, "Logistic disease incidence models and case-control studies," *Biometrika*, pp. 403–411, 1979.

[2] N. T. Bailey, ed., *The biomathematics of malaria. The Biomathematics of Diseases: 1*. 1982.

[3] J. Legrand, J. R. Egan, I. M. Hall, S. Cauchemez, S. Leach, and N. M. Ferguson, "Estimating the Location and Spatial Extent of a Covert Anthrax Release," *PLoS Comput Biol*, vol. 5, p. e1000356, Apr. 2009.

[4] A. V. Fiacco and G. P. McCormick, *Nonlinear programming: sequential unconstrained minimization techniques*. SIAM, 1990.

[5] L. M. Rios and N. V. Sahinidis, "Derivative-free optimization: a review of algorithms and comparison of software implementations," *Journal of Global Optimization*, vol. 56, pp. 1247–1293, July 2013.

[6] M. Powell, "A Survey of Numerical Methods for Unconstrained Optimization," *SIAM Review*, vol. 12, pp. 79–97, Jan. 1970.

[7] R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 2013.

[8] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability models," *Biometrics*, vol. 61, no. 3, pp. 768–769, 1965.

[9] J. MacQueen, "Some methods for classification and analysis of multivariate observations," The Regents of the University of California, 1967.

[10] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, vol. 2, pp. 559–572, Nov. 1901.

[11] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[12] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[13] J. A. Hartigan and J. A. Hartigan, *Clustering algorithms*, vol. 209. Wiley New York, 1975.

[14] P. K. Newby and K. L. Tucker, "Empirically derived eating patterns using factor or cluster analysis: a review," *Nutrition reviews*, vol. 62, no. 5, pp. 177–203, 2004.

[15] W. C. Moore, D. A. Meyers, S. E. Wenzel, W. G. Teague, H. Li, X. Li, R. D'Agostino Jr, M. Castro, D. Curran-Everett, A. M. Fitzpatrick, and others, "Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program," *American journal of respiratory and critical care medicine*, vol. 181, no. 4, pp. 315–323, 2010.

[16] A. J. Graham, P. M. Atkinson, and F. M. Danson, "Spatial analysis for epidemiology," *Acta tropica*, vol. 91, no. 3, pp. 219–225, 2004.

[17] J. A. Baecke, J. Burema, and J. E. Frijters, "A short questionnaire for the measurement of habitual physical activity in epidemiological studies.," *The American journal of clinical nutrition*, vol. 36, no. 5, pp. 936–942, 1982.

[18] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[19] S. Kotz, N. Balakrishnan, and N. L. Johnson, *Continuous multivariate distributions, models and applications*. John Wiley & Sons, 2004.

[20] D. J. Aldous, "Exchangeability and related topics," in *École d'Été de Probabilités de Saint-Flour XIIIŮ1983*, pp. 1–198, Springer, 1985.

[21] N. Dobrinkova, A. De Gaetano, T. J. R. Finnie, M. Heckel, A. Kostaridis, E. Nectarios, A. Olunczek, C. Psaroudakis, G. Seynaeve, S. Tsekeridou, and D. Vergeti, "Crisis management and disaster response tools in IMPRESS project," in *CMDR COE Proceeding 2016*, (Sofia, Bulgaria), pp. 103–124, Sept. 2016.