Applications in Industry

Kerrie K. Mengersen, Earl Duncan, Julyan Arbel, Clair Alston-Knox, Nicole White

15

Applications in Industry

Kerrie Mengersen, Earl Duncan, Julyan Arbel, Clair Alston-Knox, Nicole White

ACEMS Queensland University of Technology, Australia; ACEMS Queensland University of Technology, Australia; Laboratoire Jean Kuntzmann, Université Grenoble Alpes and INRIA Grenoble—Rhône-Alpes, France; Griffith University, Australia; Institute for Health and Biomedical Innovation, Queensland University of Technology, Australia

CONTENTS

15.1 Introduction

There are various definitions of the term *industry*, ranging from a traditional focus on manufacturing enterprises, to a slightly more relaxed inclusion of general trade, to a very broad umbrella of dedicated work. In this chapter we take a middle ground and include activities that have a commercial focus. This definition embraces an alphabet of fields, spanning agriculture, business and commerce, defence, engineering, fisheries, gas and oil, health, and so on.

A very wide range of commonly encountered problems in these industries are amenable to statistical mixture modelling and analysis. These include process monitoring or quality control, efficient resource allocation, risk assessment, prediction, and so on. Commonly articulated reasons for adopting a mixture approach include the ability to describe non-standard outcomes and processes, the potential to characterise each of a set of multiple outcomes or processes via the mixture components, the concomitant improvement in interpretability of the results, and the opportunity to make probabilistic inferences such as component membership and overall prediction.

In this chapter, we illustrate the wide diversity of applications of mixture models to

problems in industry, and the potential advantages of these approaches, through a series of case studies. The first of these focuses on the iconic and pervasive need for process monitoring, and reviews a range of mixture approaches that have been proposed to tackle complex multimodal and dynamic or online processes. The second study reports on mixture approaches to resource allocation, applied here in a spatial health context but which are applicable more generally. The next study provides a more detailed description of a multivariate Gaussian mixture approach to a biosecurity risk assessment problem, using big data in the form of satellite imagery. This is followed by a final study that again provides a detailed description of a mixture model, this time using a nonparametric formulation, for assessing an industrial impact, notably the influence of a toxic spill on soil biodiversity.

15.2 Mixtures for Monitoring

Process monitoring is an iconic problem in many industrial settings, also known as statistical process control, quality control or health monitoring (Ge et al., 2013). Historically, statistical tools developed for monitoring these processes were based on the assumption that the associated data were generated from a single distribution, representative of the population under study. It has since been acknowledged that many of these processes may be sufficiently heterogeneous to warrant the use of a mixture distribution. From their general formulation in Chapter 1, a variety of distributions can be considered as components in a mixture model, which represent interpretable features of the population or allow for flexible modelling of non-standard data. An example of such an approach, and the currency of interest in this problem, is given by Sindhu et al. (2015), who describe Bayesian estimation of Gumbel mixture models and the development of associated cumulative quality control charts.

Fault detection is another example of a common industrial problem where mixture models have been successfully applied for purposes of monitoring. Faults can be due to a range of factors, such as aging of equipment, drifting of sensors or reactions, or modifications to the underlying process (Xie & Shi, 2012). Since multiple faults may be considered simultaneously, each with different operating conditions, it is natural to consider a mixture model of some form to describe the process. A d-dimensional Gaussian mixture model is commonly assumed,

$$y_i \sim \sum_{g=1}^{G} \eta_g \mathcal{N}_d \left(\mu_g, \Sigma_g\right), \tag{15.1}$$

where y_i denotes an observed process output or a transformed output, for example, principal components. The resulting model is then combined with other statistics to construct an overall index for fault detection. An example of work in this area is provided by Yu & Quin (2008), who proposed a Gaussian finite mixture model for multimode chemical process monitoring. In this application, each mixture component described an operating mode, with the resulting model estimated by the EM algorithm. Following model estimation, the authors proposed a Bayesian approach for subsequent inference on fault detection, by first calculating posterior probabilities of component membership for each observation. These probabilities were then combined with local, component-specific Mahalanobis distances to construct an overall index for fault detection. The authors argue, through examples, that the mixture approach is superior to the more traditional multivariate process monitoring methods such as principal components and partial least squares, both of which inadequately assume the process follows a unimodal Gaussian distribution. In related work, Wen et al. (2015) propose the mixture canonical variate analysis model, in which Gaussian mixture

components are again used to describe different operating modes and singular value decomposition of the covariance matrix is employed for each cluster. Monitoring indices are then formed based on local statistics derived from the canonical variates for each cluster. These authors and others have proposed a range of variations for the use of mixture models for complex process monitoring, such as multiphase batch processes (Yu & Quin, 2009; Chen & Zhang, 2010), nonlinear multimode non-Gaussian processes (Yu, 2012a) and other dynamic or online processes (Yu, 2012b; Xie & Shi, 2012; Lin et al., 2013).

Mixtures can also be applied to a wide range of other monitoring problems. In ecology, for example, Neubauer et al. (2013) adopt a Dirichlet process mixture of multivariate Gaussians for species distribution estimation and source identification based on observed geochemical signatures. Ecological monitoring of abundance can also be framed as a binomial mixture model (Wu et al., 2015). Here, the authors cast the mixture in a Bayesian hierarchical framework in order to monitor spatially referenced replicated count data with characteristic unbalanced sampling and overdispersion. The use of mixture models to describe spatial variation in geographic monitoring has been explored by a large number of authors; one example of this is detailed in a later section of this chapter.

Estimation of species abundance is another area where mixture models offer an appealing solution. In particular, mixtures have seen wide application in dealing with zero-inflation, often encountered in abundance data (Korner-Nievergelt et al., 2015; McCarthy, 2007, pp. 264). The zero-inflated Poisson model (Lambert, 1992) is a special case of a mixture model, with components comprising of a Poisson distribution and a Dirac mass at zero, see also Chapter 9, Section 9.2.3. Membership to each component is modelled by a Bernoulli distribution with unknown probability $\eta(x_i)$, which in turn characterises the underlying zero generating process. These models take the following general form,

$$p(y_i) = \begin{cases} \eta(x_i)\,\mathbb{I}_0(y_i) + (1-\eta(x_i))\mathcal{P}(\mu(x_i)), & \text{if } y_i = 0, \\ (1-\eta(x_i))\mathcal{P}(\mu(x_i)), & \text{if } y_i > 0, \end{cases}$$

where covariate effects, x_i, are often included for predicting both the probability of a nonzero count, $\eta(x_i)$, and the Poisson rate parameter, $\mu(x_i)$. Rhodes et al. (2008) adopt a similar modelling strategy for monitoring the long term impact of chemical pollution on aquatic environments. Here, the authors formulate a zero-inflated Poisson mixed effects model to estimate the effect of copper exposure on the reproductive output of a copepod (a small crustacean found in nearly every marine and freshwater body) over three generations. In addition to covariate effects, random effects are included in each component of the proposed model, to account for correlation among observations taken from the same experimental unit (a single female copepod). In a different study, Lyashevska et al. (2016) developed a zero-inflated Poisson model with spatially correlated random effects, with an application to abundance estimation of Macoma balthica, an invertebrate found in the Wadden Sea. Similar to applications already presented, the motivation for the use of mixtures in these studies included the ability to describe the non-standard distributions of the measures considered and to make more subtle inferences. A nonparametric approach to monitoring similar industrial impacts is described in more detail in a later section of this chapter.

A final example of an industry in which monitoring is a fundamental tool is target tracking and recognition. Bayesian and non-Bayesian mixture models for clustering, classification and signal separation in this context have been used for more than a decade (see, for example, Sadjadi, 2001 and Vigneron et al., 2010) and have become more sophisticated with adaptation, realtime capabilities and ability to better distinguish foreground, background and trajectories (e.g., KaewTraKulPong & Bowden, 2002). Mixture models for human-computer interactions are also popular; for example, Pietquin (2004) discusses early

ideas for unsupervised learning of dialogue strategies, which have been refined and expanded markedly in the last ten years.

15.3 Health Resource Usage

The health industry is a very large industry, incorporating activities associated with drug development, diagnosis, surgery, rehabilitation, and other activities relating to healthcare. Good management of financial, clinical, administrative and other resources is required to provide high-quality healthcare. Therefore, any models which may assist managers of healthcare resources are a valuable tool. Two case studies are presented below which demonstrate the use of mixture models in the health industry and how they can provide management with insights into the practices regarding health resource usage.

15.3.1 Assessing the effectiveness of a measles vaccination

The goal of an immunisation program is to target certain groups of individuals with a vaccine so as to achieve herd immunity in the population. The success of an immunisation program can be assessed by studying the number of individuals who are deemed immune, as determined by high levels of antibodies, before and after administration of the vaccine. In the case of measles, individuals have traditionally been classified as immune if a serum sample indicates that their concentration of antibodies to measles is greater than some threshold. However, this approach tends to have poor test sensitivity, and does not account for different degrees of immunity.

Del Fava et al. (2012) illustrate how a mixture model can be used to estimate age-specific population prevalences and more accurately classify individuals' state of immunity. The case study in Del Fava et al. (2012) involves a national measles immunisation program in Tuscany, Italy, targeted at children less than 15 years old, conducted in 2004-05. Serum samples were collected pre- and post-immunisation (2003 and 2005-06 respectively) from individuals aged up to 49 years, and the concentration of antibodies was recorded. To account for the heterogeneity in the levels of antibody concentration, each sample of antibody concentration is modelled by a (univariate) Gaussian mixture:

$$y_i \sim \sum_{g=1}^{G} \eta_g \mathcal{N}\left(\mu_g, \sigma_g^2\right), \quad i = 1, \ldots, n,$$

where μ_g and σ_g^2 are the component-specific mean and variance, and are given non-informative priors:

$$\begin{aligned}
\mu_1 &\sim \mathcal{U}\left(y_{min}, y_{max}\right) \mathbb{I}_{(-\infty, \mu_2)}, \\
\mu_g &\sim \mathcal{U}\left(y_{min}, y_{max}\right) \mathbb{I}_{(\mu_{g-1}, \mu_{g+1})}, \quad g = 2, \ldots, G-1, \\
\mu_G &\sim \mathcal{U}\left(y_{min}, y_{max}\right) \mathbb{I}_{(\mu_{G-1}, \infty)}, \\
\sigma_g^2 &\sim \mathcal{G}^{-1}\left(0.01, 0.01\right).
\end{aligned}$$

The latent indicator is modelled by the multinomial distribution

$$z_i \sim \mathcal{M}\left(1, \eta(a_i)\right),$$

where the age-specific mixture probabilities $\eta(a_i) = (\eta_{1,a_i}, \ldots \eta_{G,a_i})$ are given the non-informative, conjugate Dirichlet prior,

$$(\eta_{1,a_i}, \ldots \eta_{G,a_i}) \sim \mathcal{D}(1, \ldots, 1), \quad i = 1, \ldots, n,$$

and a_i is the age of the i^{th} individual. The number of components is unknown, but is unlikely to be very large in this context. However, unlike the conventional dichotomous approach to classification of immunity, the number of components is unlikely to be two, even though individuals will ultimately be classified as either susceptible or immune. In the Tuscany measles immunisation program case study, separate models were fit to the pre- and post-immunisation sample data, and three and four subpopulations were identified by the models respectively. The 'optimal' number of mixture components was determined by re-fitting the models with different numbers of components, and selecting the two models with the best goodness-of-fit.

Two simplifications to this model are discussed by Del Fava et al. (2012). Firstly, the component-specific variance σ_g^2 may be considered homogeneous, i.e. $\sigma_g^2 = \sigma^2$ for all g. Secondly, the mixture probabilities $\eta = (\eta_1, \ldots \eta_G)$ can be made age-independent, following:

$$(\eta_1, \ldots \eta_G) \sim \mathcal{D}(1, \ldots, 1).$$

The results in Del Fava et al. (2012) are based on models using homogeneous variance and age-dependent mixture probabilities.

By modelling the antibody concentration using a mixture model, different levels of immunity may be observed (see Figure 15.1). For example, in the measles case study, four subpopulations were identified by the model fit to the post-immunisation data. The component with the smallest mean represented individuals who were the most susceptible and were most likely unvaccinated, while the components with the two largest means represented individuals who exhibited high degrees of immunity, where the high level of antibody concentration may be the result of the vaccination or natural infection. The remaining component represented a distinctly separate group of individuals who exhibited some degree of immunity, but with comparatively lower levels of antibody concentration.

In the case of the age-dependent model, useful inferences can be drawn from a simple time series plot of the proportion of individuals belonging to each mixture component, i.e. the prevalence, against the age of the individuals. Such a plot may help analyse the effects of initial vaccine uptake and immunisation boosts, and identify age groups which are the most susceptible and thus perhaps should be targeted in future immunisation programs.

15.3.2 Spatio-temporal disease mapping: identifying unstable trends in congenital malformations

In most disease mapping studies, the data are collected over a long period of time, but the temporal effect on the relative risk is often ignored. Abellan et al. (2008) demonstrate how a Bayesian spatio-temporal mixture model can be used to simultaneously estimate the relative risk for each area at each time point, and identify areas which exhibit a temporal pattern with "substantial and distinctive variability". Being able to detect such variability, or instability, may be of interest to medical researchers, health practitioners, and government bodies because it helps identify potential, sudden fluctuations in reported cases, either due to changes in risk factors or health care practices, necessitating further investigation.

The case study in Abellan et al. (2008) involves data on the annual number of non-chromosomal congenital malformations in England, observed across 970 artificially constructed square areas over 16 years. Due to the low rate of malformations, a binomial

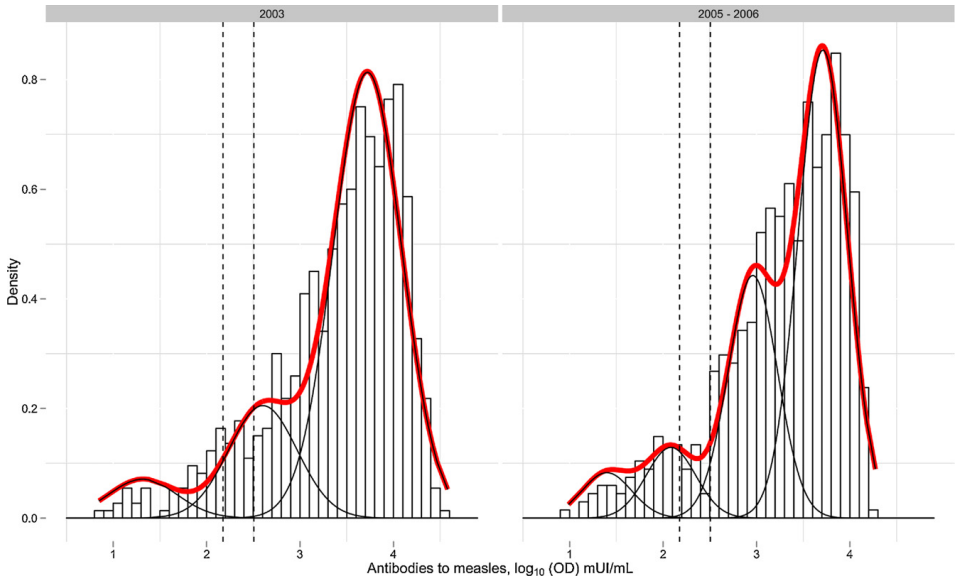


FIGURE 15.1
Histogram of the antibody concentration for the pre- and post-immunisation samples, overlaid by the estimated mixture component densities. The dashed lines represent traditional cut-off values for classifying individuals as either susceptible or immune. Reproduced with permission from Del Fava et al. (2012).

random variable is assumed to model the count data:

$$y_{it} \sim \mathcal{B}\left(n_{it}, \pi_{it}\right),$$

where n_{it} is the total number of births (including stillbirths) and π_{it} is the relative risk of congenital malformations in area $i = 1, \ldots, 970$, and year $t = 1, \ldots, 16$. The relative risk is decomposed into four parameters associated by the logistic link function,

$$\log(\pi_{it}) = \mu + \beta_i + \beta_t + \beta_{it}, \tag{15.2}$$

where μ is the estimated overall risk, β_i and β_t are the spatial and temporal random effects respectively, and β_{it} is a space-time interaction parameter. Intrinsic conditional autoregressive models are used in the specification of the priors for β_i and β_t to account for spatial and temporal heterogeneity. To account for heterogeneity of the space-time interaction, a two-component Gaussian mixture model is used,

$$\beta_{it} \sim \eta\mathcal{N}\left(0, \sigma_1^2\right) + (1-\eta)\mathcal{N}\left(0, \sigma_2^2\right),$$

where the latent indicator

$$z_{it} \sim \mathcal{M}\left(1, \eta, 1-\eta\right)$$

determines whether the variance $\sigma^2_{z_{it}}$ is small or large:

$$\sigma_1^2 \sim \mathcal{N}\left(0, 0.01\right) \mathbb{I}_{(0,+\infty)},$$
$$\sigma_2^2 \sim \mathcal{N}\left(0, 100\right) \mathbb{I}_{(0,+\infty)}.$$

Including additional covariate information in the regression equation (15.2) should be straightforward. For the full model specification, we refer the reader to Abellan et al. (2008).

By analysing the posterior probabilities that β_{it} has a large variance for a given area, one can characterise the stability of the spatial patterns of the relative risk in area i. Given the spatial nature of the data in such studies, a choropleth map may be useful for identifying whether the areas with unstable temporal trends are spatially correlated. In the congenital malformations case study, 125 (13%) of the 970 areas were identified as having unstable temporal patterns. A map of England discretised into the 970 square areas superimposed with the 125 unstable areas indicated no signs of spatial correlation.

Additional analysis of the data corresponding to the mixture component representing unstable temporal trends may provide more insight. For example, in the malformations case study, the 125 unstable areas were further classified into one of five clusters according to the similarity of their estimated space-time interaction effects. One cluster, consisting of only two areas, indicated a large spike in the risk of malformations in a particular year. Further investigations revealed that the number of kidney malformations reported in that year had increased as a result of changes to recording practices.

15.4 Pest Surveillance

Exotic and native pests in plague proportions can cause major social and economic problems in a range of industries such as agriculture and eco-tourism. The case study presented here uses Bayesian mixture models to assist managers in pre-season assignment of areas of habitat that are potentially valuable to search for fire ant infestation.

South American fire ants (*Solenopsis invicta*) were first discovered in the Port of Brisbane, Australia, in 2001. The seriousness of this threat in terms of social, environmental and economic consequences was recognised over all levels of government and affected industry, and, as a result, The National Red Imported Fire Ant Eradication Program was established in late 2001.

Since this time, data has been collected on the location of each colony that has been found, with a total of 5,027 locations spanning the years 2001–2010 being used in this analysis. However, as eradication was seen as a viable outcome, very few details of the colonies were documented or collated, resulting in the only available data being the year of discovery and the longitude and latitude of the infestation. Given this limitation, it was decided to use Landsat images taken in the winter months of each year of discovery, along with 18 potential habitat indices which may fit with the theories and observations formed by long term biological control officers within the program. The use of Landsat imagery in this analysis attempts to enhance the prospects of surveillance at a relatively low cost. As discussed in Spring & Cacho (2015), the appropriate use of remote surveillance, in combination with ground surveillance, can increase the likelihood of either containment or eradication, which is the purpose of this analysis.

15.4.1 Data and models

The aim of the analysis is to cluster areas of the Brisbane region that may be useful in identifying suitable habitat for the invasive species. An appealing starting point is to use a multivariate normal mixture model as a soft clustering technique. For computational reasons, the model was first developed without reference to spatial dependence, with the spatial component considered in a secondary analysis, following Alston et al. (2009). The

Algorithm 15.1 Gibbs sampler for estimation of parameters in a normal mixture model

1 Update $z_i \sim \mathcal{M}(1, \tau_{i1}, \tau_{i2}, \ldots, \tau_{iG})$, where

$$\tau_{ig} = \frac{\eta_g \phi(y_i|\mu_g, \Sigma_g)}{\sum_{j=1}^{G} \eta_j \phi(y_i|\mu_j, \Sigma_j)}.$$

2 Update $\eta|\mathbf{y}, \mathbf{z} \sim \mathcal{D}(e_0 + n_1, e_0 + n_2, \ldots, e_0 + n_G)$, where $n_g = \sum_{i=1}^{n} \mathbb{I}(z_i = g)$.

3 Update $\Sigma_g^{-1}|\mathbf{y}, \mathbf{z} \sim \mathcal{W}(c_g, C_g)$, where $c_g = c_0 + n_g$ and

$$C_g = C_0 + W_g + \frac{n_g N_0}{N_0 + n_g}(\bar{y}_g - b_0)(\bar{y}_g - b_0)^\top,$$

where

$$\bar{y}_g = \frac{1}{n_g}\sum_{i=1}^{n} \mathbb{I}(z_i = g) y_i, \quad W_g = \sum_{i=1}^{n} \mathbb{I}(z_i = g)(y_i - \bar{y}_g)(y_i - \bar{y}_g)^\top.$$

4 Update $\mu_g|\Sigma_g, \mathbf{y}, \mathbf{z} \sim \mathcal{N}(b_g, B_g)$, where $B_g = \Sigma_g/(N_0 + n_g)$ and

$$b_g = b_0 \cdot \frac{N_0}{N_0 + n_g} + \bar{y}_g \cdot \frac{n_g}{N_0 + n_g}.$$

non-spatial model is usually represented as in equation (15.1) which can also be represented as

$$Y_i|z_i \sim \sum_{g=1}^{G} \mathbb{I}(z_i = g) \cdot \mathcal{N}_d(y_i|\mu_g, \Sigma_g),$$

where $\mathcal{N}_d$ denotes a multivariate normal density with mean μ_g and variance-covariance matrix Σ_g. As described in Chapter 1, a vector $\mathbf{z} = \{z_1, z_2, \ldots, z_n\}$ can be associated with this sample. It consists of a set of unobserved indicator vectors denoting component membership, which are estimated as another parameter in the model using Gibbs samplimg. As in earlier chapters, η_g represents the weight (proportion) of each component in the model and determines the prior probability $\mathrm{P}(z_i = g|\eta_g) = \eta_g$. Denote $\eta = (\eta_1, \ldots, \eta_G)$.

We use standard conjugate priors in this mixture model. Specifically,

$$\begin{aligned} \Sigma_g^{-1} &\sim \mathcal{W}(c_0, C_0), \\ \mu_g|\Sigma_g &\sim \mathcal{N}\left(b_0, \frac{1}{N_0}\Sigma_g\right), \\ \eta &\sim \mathcal{D}_G\left(e_0\right). \end{aligned}$$

Frühwirth-Schnatter (2006, p.192-193) provides an extensive discussion on the choice of hyperparameters for these prior distributions. In this analysis, we have followed the recommendations of Robert (1995) for hyperparameter selection. The resulting algorithm for estimating the parameters of this mixture model is well known. It is a two stage Gibbs sampler, outlined in Algorithm 15.1.

As the number of pixels in an image was quite large, in excess of 2 million, to assist our clients implement these models, a library was developed in PyMCMC (Strickland et al., 2012), which includes a Gibbs sampler to estimate the mixture model parameters, as per this algorithm. In this analysis, Step 1 of the algorithm is the most laborious, and as such

Algorithm 15.2 Label switching re-ordering scheme used in Gibbs sampler for estimation of parameters of the normal mixture model.

1 Perform mixture model updating for a specified burn-in period.
2 During the next 10,000 iterations, calculate the likelihood of the mixture at every 100th iteration, based on the current parameter estimates.
3 From these calculated likelihoods, choose the set of parameters that maximises the likelihood. This parameter set then becomes the base ordering, from which any label switching is detected. Note, there is no mean or variance ordering imposed on this set.
4 For each iteration thereafter, test the updated parameters for the latent variable against this base ordering. The possibility of switched labels is then determined. This is computed by comparing the current allocations of $\mathbf{z}$ to the allocations of $\mathbf{z}$ during the chosen optimal iteration. A misclassification matrix, C, is constructed, and the cost of misclassifications determined using the Munkres algorithm (also called the Hungarian algorithm or the Kuhn-Munkres algorithm, see Kuhn, 1955).

was programmed with the option of parallel computation for the end user. Additionally, rather than store all the iterates, we wished to compute posterior estimates "on the run", saving on storage space. To enable this, we needed to deal with label switching within the mixture, and this was achieved using Algorithm 15.2, which is a single chain variation on that presented by Cron & West (2011).

Technically, there is no limit on the number of dimensions (variables) we could fit in this mixture model. So potentially, we could include all 18 variables in the analysis. However, from a pragmatic viewpoint, there is little to be gained, and a large computational burden to incur, by fitting variables that have no influence on habitat suitability. For this reason, we decided to pre-determine which of the 18 variables were modelled via mixtures. Using expert opinion and CART analysis, it was decided to use Landsat band 3 (visible red), Landsat band 6 (mid infrared) and a soil brightness index to assess the probability that the area associated with each pixel is habitable terrain for inclusion in the upcoming surveillance season. The multivariate analysis allowed managers to create meaningful clusterings that reflect the sometimes complex combinations of conditions that form habitat suitability, rather than relying on single derived indices. Figure 15.2 illustrates the difference in these three variables between known infestation sites and non-infested areas.

15.4.2 Resulting clusters

To test the potential of the model, we estimated the parameters of a multivariate normal mixture model on a Landsat image from 2011 based on Landsat bands 3 and 6, and the soil brightness index. The component estimates for the component weights, means, and covariance matrices are given in Table 15.1. As the density estimates in Figure 15.2 indicate that positive centered values of these pixels are dominant in regions of known *S. invicta* colonies, it can be seen that components 3 to 6 are of most interest for suitable habitat, component 2 is of some interest, as each mean is positive but the distribution straddles the negative valued regions, and we would consider areas of reasonable size which are allocated to this component. Component 1, which has negative valued means, is considered not suitable for *S. invicta* habitat in terms of surveillance, due to the initial findings of the CART analysis. Component 1 is associated with over 85% of the area in the Landsat image, so this is helpful in terms of surveillance planning in that the tracts of land associated with component 1 can be largely discounted for surveillance if budgetary restraints are an issue.

A big advantage of viewing the region in terms of this mixture model, from a management

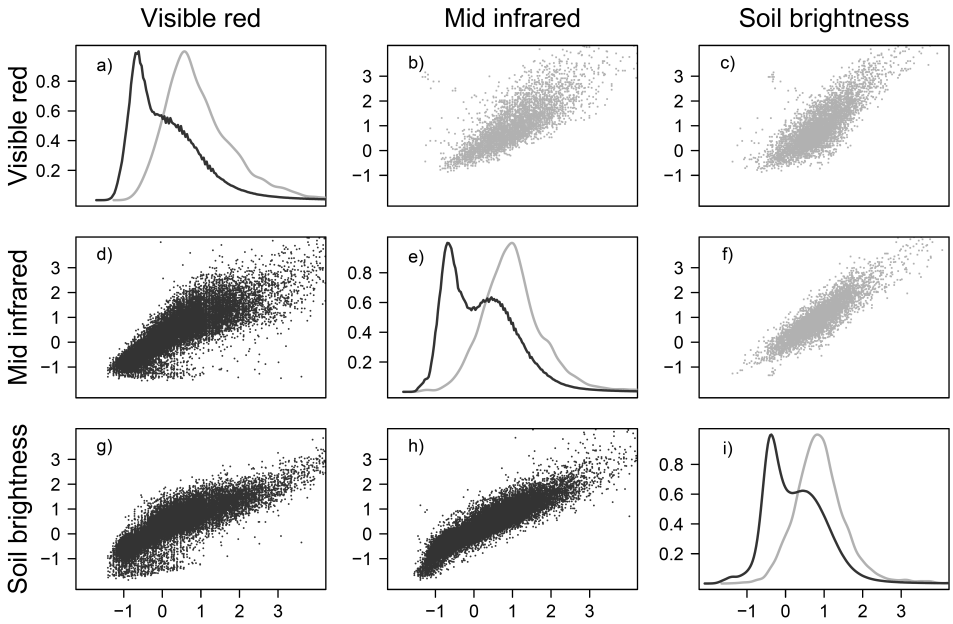


FIGURE 15.2
Density and scatter plots of the three variables of interest in habitat modelling. a) Density of visible red in areas of previously detected fire ant colonies (light gray) and areas not currently infested (dark gray). b) Scatter plot of visible red against mid infrared in areas of known infestation. c) Scatter plot of visible red against soil brightness in areas of known infestation. d) Scatter plot of visible red against mid infrared in non-infested areas. e) Density of mid infrared in areas of previously detected fire ant colonies (light gray) and areas not currently infested (dark gray). f) Scatter plot of mid infrared against soil brightness in areas of known infestation. g) Scatter plot of visible red against soil brightness in non-infested areas. h) Scatter plot of mid infrared against soil brightness in non-infested areas. i) Density of soil brightness in areas of previously detected fire ant colonies (light gray) and areas not currently infested (dark gray).

point of view, is the ability to create posterior probability maps which will predict the likelihood of areas belonging to suitable habitat, and combining this with expert knowledge from the ground workers, previous finds and budgetary restraints to make effective decisions in terms of where to concentrate the surveillance effort in any year. Figure 15.3 illustrates the range of estimated posterior probabilities for each pixel belonging to components 2 to 6. These values are converted into a map within ArcGIS software and layered with other relevant input to form a surveillance strategy. The probability aids managers to make decisions about prioritising areas.

Generally, we found that this method highlighted new areas of urban growth, which anecdotally were developments where experienced staff expected the possibility of fire ant infestations. Fire ant colonies that were missed using this technique tended to be roadside infestations caused by transportation via tyres, and realistically, these are not habitats in which we would wish to perform surveillance in, relying instead on public education to detect infestations within habitats best defined by component 1 of the mixture.

TABLE 15.1
Estimated component means, variances and weights from the mixture model for Landsat band 3 (LS_3), Landsat band 6 (LS_6) and Soil Brightness (SB). Associated credible intervals given in brackets.

Cluster	Variable	Mean	Covariance matrix		
1	LS_3	-0.305	26.80		
		(-0.307,-0.303)	(26.6, 27.0)		
	LS_6	-0.260	-25.50	34.5	
		(-0.262, -0.258)	(-25.7, -25.4)	(34.4, 34.7)	
	SB	-0.193	7.57	-14.7	9.02
		(-0.195, -0.191)	(7.52, 7.62)	(-14.8, -14.6)	(8.97, 9.06)
Weight	0.8510				
	(0.8490, 0.8520)				
2	LS_3	1.35	7.83		
		(1.34, 1.36)	(7.71, 7.95)		
	LS_6	1.37	0.92	5.99	
		(1.36, 1.37)	(0.82, 1.01)	(5.91, 6.06)	
	SB	1.01	-8.59	-6.18	19.10
		(1.00, 1.02)	(-8.75, -8.44)	(-6.30, -6.07)	(18.80, 19.30)
Weight	0.1030				
	(0.1020, 0.1040)				
3	LS_3	1.85	9.5		
		(1.80, 1.90)	(9.20, 9.82)		
	LS_6	0.978	3.3	8.64	
		(0.920, 1.04)	(3.05, 3.56)	(8.17, 9.08)	
	SB	0.597	-11.3	-11.5	21.7
		(0.560, 0.638)	(-11.8, -10.8)	(-12.0, -11.1)	(20.9, 22.4)
Weight	0.0308				
	(0.0302, 0.0314)				
4	LS_3	2.62	3.630		
		(2.51, 2.74)	(3.29, 3.99)		
	LS_6	2.92	0.95	0.96	
		(2.85, 2.98)	(0.75, 1.15)	(0.81, 1.11)	
	SB	2.58	-4.66	-1.15	8.54
		(2.51, 2.65)	(-5.13, -4.22)	(-1.40, -0.89)	(7.62, 9.50)
Weight	0.0009				
	(0.0008, 0.0010)				
5	LS_3	3.92	12.8		
		(3.82, 4.03)	(12.0, 13.8)		
	LS_6	3.19	5.78	4.24	
		(3.13, 3.25)	(5.24, 6.28)	(3.85, 4.64)	
	SB	2.46	-20.5	-10.6	34.7
		(2.38, 2.55)	(-22.0, -19.1)	(-11.5, -9.6)	(32.1, 37.4)
Weight	0.0132				
	(0.0119, 0.0144)				
6	LS_3	8.08	0.73		
		(7.97, 8.19)	(0.66, 0.80)		
	LS_6	4.83	0.41	0.95	
		(4.74, 4.91)	(0.37, 0.45)	(0.88, 1.02)	
	SB	5.71	-1.29	-1.45	3.23
		(5.55, 5.86)	(-1.40, -1.19)	(-1.56, -1.34)	(2.94, 3.50)
Weight	0.0012				
	(0.0011, 0.0013)				

15.5 Toxic spills

Toxic spills are externalities of industrial activity that affect soils. Their impact on the biodiversity is one of many detrimental consequences of toxic spills on an ecosystem. Bio-

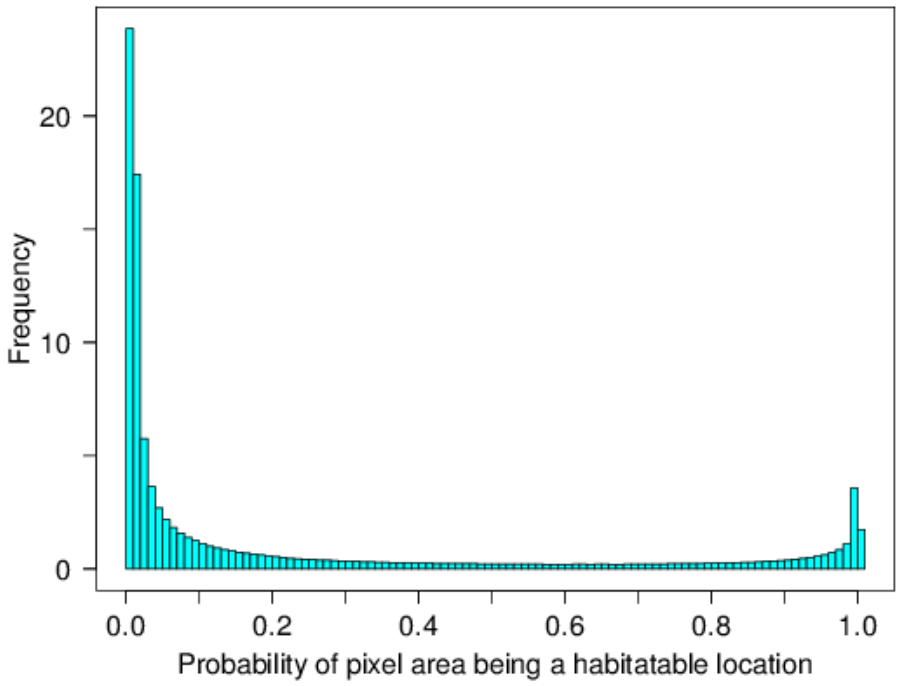


FIGURE 15.3
Histogram representing posterior probability of individual pixels in 2011 Landsat image belonging to habitat suitable for fire ant infestation surveillance.

diversity is a measure of the variety of organisms present in an ecosystem. Since mixture models are probability models for representing the presence of subpopulations within an overall population, it is reasonable to use mixtures for modelling biodiversity. We will pursue this approach here, under a Bayesian nonparametric (BNP) framework, see Chapter 6 for a comprehensive review of BNP mixture models.

The data of the present case study consist of microbial communities, or groups of species, observed as counts at locations in the soil, or sites, along with a toxic contaminant measurement. The composition of species may differ among the sites, and the main inferential interest amounts to understanding the contaminant impact on species diversity.

In contrast with most of the mixture models presented in this handbook, the identity of mixture components is actually observed in the data. As we shall see, this feature leads to inferential and computational techniques contrasting with other approaches.

Note that the exposition is in terms of species due to the actual application to toxic spills. Nevertheless, the approach is not limited to species sampling problems. This application is based on Arbel et al. (2015, 2016).

15.5.1 Data and model

The data consist of a soil microbial data set acquired across a hydrocarbon contamination gradient at the location of a fuel spill at Australia's Casey Station in East Antarctica (110° 32′ E, 66° 17′ S), along a transect at 22 locations. Microbes are classified as Operational

Taxonomic Units, that we also generically refer to as species. Species measurements are paired with a contaminant called Total Petroleum Hydrocarbon (TPH), suspected to impact diversity. We refer to Snape et al. (2015) for a complete account on the data set acquisition.

The state space is the set of species. We label them with positive integers, and order them in overall (over all sites) decreasing abundance. Probabilities of presence, say η_g for species g, are a set of self-evident parameters in this kind of species sampling models. They are positive and sum up to one. When the total number of species is fixed, say G, the presence probabilities are element of the unit simplex of dimension $G-1$. Numerous community summaries of interest to ecologists are described in terms of probabilities of presence, such as diversity, richness, evenness, to name just a few. Instances of predominant indices in ecology include the Shannon index $-\sum_g \eta_g \log \eta_g$ (or entropy), the Simpson index (or Gini index) $1-\sum_g \eta_g^2$, and the Good index $-\sum_g \eta_g^\alpha (\log \eta_g)^\beta$, $\alpha, \beta \geq 0$, which generalises both.

The data can be further described as follows. To each site $i = 1, \ldots, I$ corresponds a covariate value $x_i \in \mathcal{X}$, where the space $\mathcal{X}$ is a subset of $\mathbb{R}$, for instance $[0, \infty)$ in the present case. Individual observations $y_{n,i}$ at site i are indexed by $n = 1, \ldots, n_i$, where n_i denotes the total abundance, or number of observations, at site i. Observations $y_{n,i}$ take on positive natural numbers $g \in \{1, \ldots, G_i\}$ where G_i denotes the number of distinct species observed at site i. We denote by $(\mathbf{x}, \mathbf{y})$ the observations over all sites, where $\mathbf{x} = (x_i)_{i=1,\ldots,I}$, $\mathbf{y} = (y_i^{n_i})_{i=1,\ldots,I}$ and $y_i^{n_i} = (y_{n,i})_{n=1,\ldots,n_i}$. The abundance of species g at site i is denoted by n_{ig}, representing the number of times that $y_{n,i} = g$ with respect to index n. The relative abundance satisfies $\sum_{g=1}^{G_i} n_{ig} = n_i$.

The multinomial distribution provides a natural framework when the sampling process consists of independent and identically distributed observations of a fixed number of species, say G (see applications in ecology like Fordyce et al., 2011; De'ath, 2012; Holmes et al., 2012). Under this framework, individual observations follow a categorical distribution, which is a generalisation of the Bernoulli distribution when the sample space is a set of G, greater than two, items. Namely, the probability mass function of an individual observation y_n, which can take on a value of a species in $\{1, \ldots, G\}$, is $\mathrm{P}(y_n = g) = \eta_g$. A tantamount notation, also more reminiscent of mixture models, is

$$y_n \overset{\text{ind}}{\sim} \sum_{g=1}^{G} \eta_g \delta_g, \tag{15.3}$$

where δ_g denotes a Dirac point mass at g. We shall adopt an extension of model (15.3) in two respects. First, we will not assume that the total number of species G is known. To this aim, we let the number of species take arbitrarily large values, hence we replace the finite sum in (15.3) by a countable infinite sum, leading us to adopt a nonparametric approach. As a consequence, the weights $(\eta_1, \eta_2, \ldots)$ become infinite vectors. Their elements sum up to one and belong to the simplex of infinite dimension. Second, due to the site-by-site nature of the data, we formulate site-wise independent, but not identically distributed, generative models. Each site i is parameterised by a vector of presence probabilities denoted by $\eta(x_i) = (\eta_1(x_i), \eta_2(x_i), \ldots)$, and we denote by $\eta = (\eta(x_1), \ldots, \eta(x_I))$ the full parameter. Taking into account these desiderata, we end up with the following data model which is a mixture model with a countable infinite number of components

$$y_{n,i} \,|\, \eta(x_i), x_i \overset{\text{ind}}{\sim} \sum_{g=1}^{\infty} \eta_g(x_i) \delta_g, \tag{15.4}$$

for $i = 1, \ldots, I$, $n = 1, \ldots, n_i$. We assume a fixed design, meaning that the covariates x_i are not randomized. Additionally, we adhere to a Bayesian viewpoint and need to endow the parameter η with a prior distribution.

The initial motivation for the prior distribution on the presence probabilities η stems from dependent Dirichlet processes introduced by MacEachern (1999). The Dirichlet process (Ferguson, 1973) is a popular Bayesian nonparametric distribution for species modelling which conveys an interesting natural clustering mechanism. Dependent Dirichlet processes were proposed by MacEachern in order to extend Dirichlet processes to multiple-site situations, and to allow for borrowing of strength across the sites, see also Chapter 6, Section 6.2.3. Dirichlet processes (and their dependent version) are (almost surely) discrete random probability measures, hence they consist of random weights and random locations. Of interest for us is the distribution of the random weights which shall constitute the prior distribution on the presence probabilities. We first describe the distribution of the weights of the Dirichlet process, also known as the Griffiths–Engen–McCloskey distribution (GEM), used as a prior for $\eta(x)$ for any given covariate value x, and then turn to describe the distribution of the weights of the dependent Dirichlet process, that we call dependent GEM, used as a joint prior for $\eta = (\eta(x_1), \ldots, \eta(x_I))$.

The marginal prior distribution on $\eta(x)$, marginal meaning at a given covariate value x, is defined in a constructive way, called *stick-breaking*, in the following way (Sethuraman, 1994). Let us introduce i.i.d. Beta random variables

$$V_g(x) \overset{iid}{\sim} \mathcal{B}e(1, \alpha), \quad g \geq 1,$$

where $\alpha > 0$. Then the prior distribution induced on $\eta(x)$ by setting $\eta_1(x) = V_1(x)$ and, for $g > 1$,

$$\eta_g(x) = V_g(x) \prod_{l<g} (1 - V_l(x)), \tag{15.5}$$

is called the Griffiths–Engen–McCloskey distribution, and α is called the precision parameter.

For an exhaustive description of the prior distribution on η, the marginal description (15.5) needs be complemented by specifying a distribution for stochastic processes $(V_g(x), x \in \mathcal{X})$, for any positive integer g. Since (15.5) requires i.i.d. Beta marginals, natural candidates are i.i.d. Beta processes. A simple yet effective construct to obtain a Beta process is to transform a Gaussian process by the inverse cumulative distribution function (cdf) transform as follows. Denote by $z \sim \mathcal{N}(0, \sigma^2)$ a Gaussian random variable, by Φ_σ its cdf and by F_α the cdf of a $\mathcal{B}e(1, \alpha)$ random variable. Then $F_\alpha^{-1}(u) = 1 - (1 - u)^{1/\alpha}$, and the random variable

$$V = h_{\sigma,\alpha}(z) = F_\alpha^{-1} \circ \Phi_\sigma(z) \tag{15.6}$$

is $\mathcal{B}e(1, \alpha)$ distributed. The idea of including a transformed Gaussian process within a stick-breaking process is used in previous articles (see for instance Rodriguez & Dunson, 2011).

Of the full path of a Gaussian process on $\mathcal{X}$, only the values at observed covariates $\mathbf{x} = (x_1, \ldots, x_I)$ are used. For any positive integer g, denote these values by $z_g = (z_g(x_1), \ldots, z_g(x_I))$, and let $\mathbf{z} = (z_1, z_2, \ldots)$ be a set of i.i.d. copies. Then the inverse cdf transform (15.6) maps $\mathbf{z}$ on a set of i.i.d. copies of Beta vectors $\mathrm{V} = (\mathrm{V}_1, \mathrm{V}_2, \ldots)$. In turn, the stick-breaking construction (15.5) maps V on the set of presence probabilities $\eta = (\eta_1, \eta_2, \ldots)$. Hence, the prior distribution on the presence probabilities η is the distribution induced under the composition of transforms (15.6) and (15.5). We denote it by

$$\eta \sim \text{Dep-GEM}(\alpha, \lambda, \sigma^2), \tag{15.7}$$

where λ and σ^2 are two parameters of the Gaussian processes that we describe now.

A Gaussian process is fully specified by a mean function, which we take equal to zero, and a covariance function Σ defined by

$$\Sigma(x_i, x_l) = \mathrm{Cov}\big(z_g(x_i), z_g(x_l)\big). \tag{15.8}$$

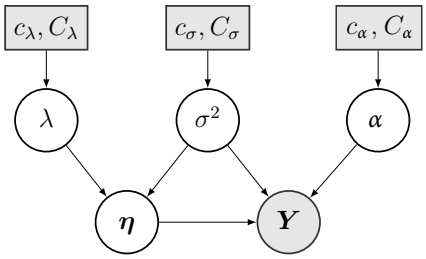
We control the overall variance of z_g by a positive pre-factor σ^2 and write $\Sigma = \sigma^2\tilde{\Sigma}$ where $\tilde{\Sigma}$ is normalised in the sense that $\tilde{\Sigma}(x_i, x_i) = 1$ for all i. Possible choices of covariance structures include the squared exponential, Ornstein–Uhlenbeck, and rational quadratic covariance functions, see Rasmussen & Williams (2006) for more details. We work equally with one of these three options, without trying to learn the covariance structure, but only a parameter λ involved in all three called the length-scale of the process. It tunes how far apart two points in the space $\mathcal{X}$ have to be for the process to change significantly. The shorter λ is, the rougher are the paths of the process. We stress the dependence on λ by denoting $\Sigma = \sigma^2\tilde{\Sigma}_\lambda$. The prior distribution of z_g is the following multivariate normal

$$p(z_g|\mathbf{x}, \sigma^2, \lambda) \propto \big(\sigma^{2I}|\tilde{\Sigma}_\lambda|\big)^{-1/2} \exp\big(-\tfrac{1}{2\sigma^2} z_g^\top \tilde{\Sigma}_\lambda^{-1} z_g\big).$$

The prior distribution is complemented by specifying the prior distribution $p(\sigma^2, \lambda, \alpha)$ over the hyperparameters. We adopt independent prior distributions, namely $p(\sigma^2, \lambda, \alpha)$ takes the form $p(\sigma^2)p(\lambda)p(\alpha)$. More specifically Gamma prior distributions are used for all three hyperparameters: the inverse variance $1/\sigma^2$, the inverse length-scale $1/\lambda$ and the precision parameter α. These prior distributions are also common choices in the absence of dependence since they turn out to be conjugate.

The complete Bayesian model is given by the following mixture model:

$$\begin{aligned}
y_{n,i} \,|\, \eta(x_i), x_i &\overset{\text{ind}}{\sim} \sum_{g=1}^{\infty} \eta_g(x_i)\delta_g, \\
i &= 1, \ldots, I,\ n = 1, \ldots, n_i, \\
\eta &\sim \text{Dep-GEM}(\alpha, \lambda, \sigma^2), \\
\sigma^2 &\sim \mathcal{IG}(c_\sigma, C_\sigma), \\
\lambda &\sim \mathcal{IG}(c_\lambda, C_\lambda), \\
\alpha &\sim \mathcal{G}(c_\alpha, C_\alpha),
\end{aligned}$$



with fixed shape parameters $c_\sigma, c_\lambda, c_\alpha$ and scale parameters $C_\sigma, C_\lambda, C_\alpha$. In the graphical representation of the model, rectangles indicate fixed parameters, circles indicate random variables, and filled-in shapes indicate known values.

15.5.2 Posterior sampling and summaries

Here we expose posterior sampling in terms of the Gaussian processes $\mathbf{z}$, keeping in mind that the main parameters of interest, the presence probabilities η, can be recovered through the composition of the inverse cdf transform (15.6) and the stick-breaking construction (15.5). The likelihood can be easily obtained from these two transforms as

$$p(\mathbf{y}|\mathbf{z}, \mathbf{x}, \sigma^2, \alpha) = \prod_{g=1}^{G}\prod_{i=1}^{I} h_{\sigma,\alpha}(z_g(x_i))^{n_{ig}}(1 - h_{\sigma,\alpha}(z_g(x_i)))^{\bar{n}_{i,g+1}},$$

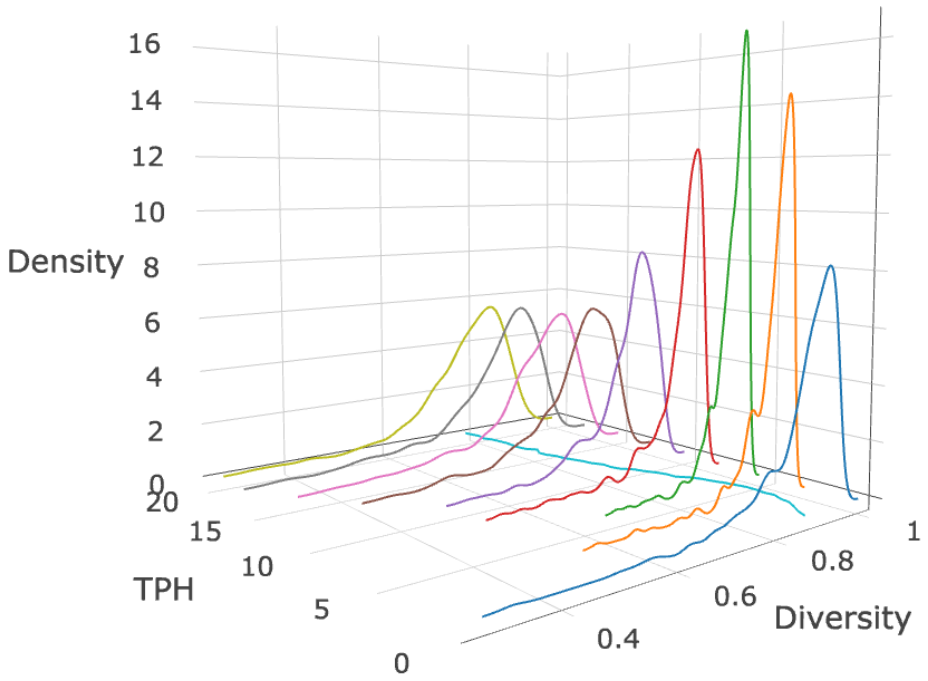


FIGURE 15.4
Posterior distributions of diversity at varying pollution levels from 0 to 20 TPH units, along with the posterior mean of the diversity.

where we denote by $\bar{n}_{i,g+1} = \sum_{l>g} n_{il}$ the sum of abundances of species $\{g+1, g+2, \ldots\}$ at site i. The posterior distribution is then

$$p(\mathbf{z}, \lambda, \sigma^2, \alpha | \mathbf{y}, \mathbf{x}) \propto p(\mathbf{y}|\mathbf{z}, \mathbf{x}, \sigma^2, \alpha) p(\mathbf{z}|\mathbf{x}, \sigma^2, \lambda) p(\sigma^2) p(\lambda) p(\alpha),$$

where $p(\mathbf{z}|\mathbf{x}, \sigma^2, \lambda) = \prod_{g=1}^{G} p(z_g|\mathbf{x}, \sigma^2, \lambda)$.

Sampling from the posterior distribution of $(\mathbf{z}, \sigma^2, \lambda, \alpha)$ in the Dep-GEM model is performed by a Markov chain Monte Carlo algorithm comprising Gibbs and Metropolis–Hastings steps. It proceeds by sequentially updating each parameter $\mathbf{z}$, σ^2, λ and α via its conditional distribution as follows.

(a) Conditional for $\mathbf{z}$: we use G independent Metropolis algorithms with Gaussian proposals. The covariance matrix for the proposal is set proportional to the prior covariance matrix $\tilde{\Sigma}_\lambda$. For any $g \in \{1, \ldots, G\}$, the target distribution is

$$p(z_g|\cdot) \propto p(\mathbf{y}|\mathbf{z}, \mathbf{x}, \sigma^2, \alpha) p(z_g|\mathbf{x}, \sigma^2, \lambda).$$

(b) Conditional for σ^2: Metropolis–Hastings algorithm with a Gaussian proposal left-truncated to 0 with target distribution

$$p(\sigma^2|\cdot) \propto p(\mathbf{y}|\mathbf{z}, \mathbf{x}, \sigma^2, \alpha) p(\mathbf{z}|\mathbf{x}, \sigma^2, \lambda) p(\sigma^2).$$

(c) Conditional for λ: Metropolis–Hastings algorithm with a Gaussian proposal left-truncated to 0 with target distribution

$$p(\lambda|\cdot) \propto p(\mathbf{z}|\mathbf{x}, \sigma^2, \lambda) p(\lambda).$$

(d) Conditional for α: Metropolis algorithm with a Gaussian proposal left-truncated to 0 with target distribution

$$p(\alpha|\cdot) \propto p(\mathbf{y}|\mathbf{z}, \mathbf{x}, \sigma^2, \alpha)p(\alpha).$$

The posterior sample allows for a probabilistic evaluation of various quantities of interest for ecologists. Figure 15.4 provides posterior distributions of diversity at varying pollution levels from 0 to 20 TPH units. This shows a better posterior precision around 5 TPH units. The posterior mean of the diversity first increases with the pollution level with a maximum at 4 TPH units, and then decreases. Such a variation may depict a hormetic effect, a dose-response phenomenon characterised by favorable responses to low exposures to pollutant. Additionally to diversity, so-called *effective concentrations* are highly relevant criteria in determining guidelines for protection of an ecosystem. The effective concentration at level ℓ, denoted by EC_ℓ, is the concentration of a contaminant that causes $\ell\%$ effect on the population relative to the baseline community (e.g. Newman, 2012). For example, the EC_{50} is the median effective concentration and represents the concentration of a contaminant which induces a response halfway between the control baseline and the maximum after a specified exposure time. Estimation of both diversity and effective concentrations is straightforward from a posterior sample. See Arbel et al. (2015) for detailed results.

15.6 Concluding Remarks

This chapter has showcased a number of industrial applications where mixture modelling has been shown to be a powerful tool for inference. Whilst the applications presented were diverse, spanning manufacturing, ecology and health, each shared the common motivation of characterising heterogeneity by multiple components, with each component describing a key feature of the population or process under study. The breadth of applications also demonstrated the flexibility of mixture specification, with models varying in terms of distributional assumptions for individual components, the form of the mixing proportions and the use of mixtures as generative models for the observed data versus as a prior distribution for unknown parameters.

A review of monitoring studies strengthened the case for mixture models as an elegant solution for explaining observed heterogeneity, in cases where the use of a single distribution is inadequate for reliable inference. In studies of fault detection, the utility of mixture models was two-fold: (i) as a way of characterising different operating modes of a process; and (ii) to construct a meaningful, global summary of the process, taking into account component membership uncertainty. In other examples presented, focused in ecology, the flexibility of mixtures in terms of component specification was highlighted. This flexibility extended to different types of data (continuous, count) and mixture component specification to account for zero inflation, for improved predictions of species abundance.

The two cases studies pertaining to health resource usage showed how mixture models can account for heterogeneity relating to tangible characteristics, like antibody concentration, and technical constructs, such as a random effect for space-time interaction. This ability to address different sources of heterogeneity is a result of applying the mixture model in different ways. In the first case study, the likelihood was represented by the mixture model directly, as is typical in many applications. In the second case study, however, the mixture model was applied to a random effect parameter, or more specifically, its prior distribution. In both case studies, the mixture models were able to identify emerging patterns and guide the management of health resources.

In the pest surveillance example, the multivariate Bayesian mixture model when applied to data drawn from standard Landsat images, is shown to be an effective method of estimating the probability of land in an urban space being suitable habitat for an invasive pest. Additionally, these models can be used to either cluster image pixels into either a most probable component or the probability of a pixel being suitable habitat, if suitability is encompassed by several clusters. The model flexibility, combined with the Bayesian approach using MCMC, allows researchers and managers to value add and extend scenario testing with relative ease.

The flexibility of the Bayesian mixture model is likely to make this modelling procedure extensible to other types of landuse (for instance, rural) and other forms of imagery, such as aerial and drone captured snapshots, ASTER and Spot 5. This is an active area of research globally, as technology becomes more readily available. Estimation of bare earth is interesting in many applications on a worldwide scale, with reports such as Weber et al. (2010) indicating that this research is ongoing in the quest to detect this land use pattern using remote sensing data.

The last case study instantiated how infinite mixtures and Bayesian nonparametric methods can be successfully applied to industry related issues, namely the influence of a toxic spill on soil biodiversity. Mixtures were demonstrated to adequately model covariate-dependent species data, allowing one to make useful inference on biodiversity, effective concentrations, etc, as well as to provide predictions outside the range of observed covariates.

Bibliography

Abellan, J. J., Richardson, S. & Best, N. (2008). Use of space-time models to investigate the stability of patterns of disease. *Environmental Health Perspectives* **116**, 1111–1119.

Alston, C. L., Mengersen, K. L. & Gardner, G. E. (2009). A new method for calculating the volume of primary tissue types in live sheep using computed tomography scanning. *Animal Production Science* **49**, 1035–1042.

Arbel, J., Mengersen, K. L., Raymond, B., Winsley, T. & King, C. (2015). Application of a Bayesian nonparametric model to derive toxicity estimates based on the response of antarctic microbial communities to fuel contaminated soil. *Ecology and Evolution* **5**, 2633–2645.

Arbel, J., Mengersen, K. L. & Rousseau, J. (2016). Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. *Annals of Applied Statistics* **10**, 1496–1516.

Chen, T. & Zhang, J. (2010). On-line multivariate statistical monitoring of batch processes using Gaussian mixture model. *Computers and Chemical Engineering* **34**, 500–507.

Cron, A. J. & West, M. (2011). Efficient classification-based relabeling in mixture models. *The American Statistician* **65**, 16–20.

De'ath, G. (2012). The multinomial diversity model: linking Shannon diversity to multiple predictors. *Ecology* **93**, 2286–2296.

Del Fava, E., Shkedy, Z., Bechini, A., P., B. & P., M. (2012). Towards measles elimination in Italy: Monitoring herd immunity by Bayesian mixture modelling of serological data. *Epidemics* **4**, 124–131.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Fordyce, J. A., Gompert, Z., Forister, M. L. & Nice, C. C. (2011). A hierarchical Bayesian approach to ecological count data: a flexible tool for ecologists. *PLoS ONE* **6**, e26785.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer-Verlag, New York.

Ge, Z., Song, Z. & Gao, F. (2013). Review of recent research on data-based process monitoring. *Industrial Engineering and Chemical Research* **52**, 3542–3562.

Holmes, I., Harris, K. & Quince, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PloS One* **7**, e30126.

KaewTraKulPong, P. & Bowden, R. (2002). An improved adaptive background mixture model for real-time tracking with shadow detection. In *2nd European Workshop on Advanced Video Based Surveillance Systems, 2001*.

Korner-Nievergelt, F., Roth, T., von Felten, S., Guélat, J., Almasi, B. & Korner-Nievergelt, P. (2015). *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS and STAN*. London: Academic Press.

Kuhn, H. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **2**, 83–97.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

Lin, Y., Chen, M. & Zhou, D. (2013). Online probabilistic operational safety assessment of multi-mode engineering systems using Bayesian methods. *Reliability Engineering and System Safety* **119**, 150–157.

Lyashevska, O., Brus, D. & Meer, J. (2016). Mapping species abundance by a spatial zero-inflated Poisson model: a case study in the Wadden Sea, the Netherlands. *Ecology and Evolution* **6**, 532–543.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*. Alexandria, VA: American Statistical Association, pp. 50–55.

McCarthy, M. A. (2007). *Bayesian Methods for Ecology*. London: Cambridge University Press.

Neubauer, P., Shima, J. S. & Swearer, S. E. (2013). Inferring dispersal and migrations from incomplete geochemical baselines: analysis of population structure using Bayesian infinite mixture models. *Methods in Ecology and Evolution* **4**, 836–845.

Newman, M. C. (2012). *Quantitative Ecotoxicology*. CRC Press.

Pietquin, O. (2004). *A Framework for Unsupervised Learning of Dialogue Strategies*. Ph.D. thesis, Faculté Polytechnique de Mons, TCTS Lab, Belgique.

Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rhodes, J. R., Grist, E. P. M., Kwok, K. W. H. & Leung, K. M. Y. (2008). A Bayesian mixture model for estimating intergeneration chronic toxicity. *Environmental Science and Technology* **42**, 8108–8114.

Robert, C. (1995). Mixtures of distributions: Inference and estimation. In *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson & D. Spiegelhalter, eds. Chapman & Hall/CRC Interdisciplinary Statistics.

Rodriguez, A. & Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6**, 145–177.

Sadjadi, F. A., ed. (2001). *Automated Target Recognition*, vol. 11 of *SPIE Conference Volume*, Bellingham, WA, USA. SPIE.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Sindhu, T. N., Riaz, M., Aslam, M. & Ahmed, Z. (2015). Bayes estimation of Gumbel mixture models with industrial applications. *Transactions of the Institute of Measurement and Control* **38**, 201–214.

Snape, I., Siciliano, S. D., Winsley, T., van Dorst, J., Mukan, J., Palmer, A. S. & Lagerewskij, G. (2015). *Operational Taxonomic Unit (OTU) Microbial Ecotoxicology data from Macquarie Island and Casey Station: TPH, Chemistry and OTU Abundance data*. Australian Antarctic Data Centre.

Spring, D. & Cacho, O. J. (2015). Estimating eradication probabilities and trade-offs for decision analysis in invasive species eradication programs. *Biological Invasions* **17**, 191–204.

Strickland, C. M., Denham, R. J., Alston, C. L. & Mengersen, K. L. (2012). A Python package for Bayesian estimation using Markov Chain Monte Carlo. In *Case Studies in Bayesian Statistical Modelling and Analysis*, C. L. Alston, K. L. Mengersen & A. N. Pettitt, eds. John Wiley & Sons, Ltd, pp. 421–460.

Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R. & Vincent, E., eds. (2010). *Latent Variable Analysis and Signal Separation: 9th International Conference, LVA/ICA 2010, St. Malo, France, September 27-30, 2010*, vol. 6365. Springer.

Weber, K., Glenn, N. & Tibbitts, J. (2010). Investigation of potential bare ground modeling techniques using multispectral satellite imagery. Tech. Rep. NNG06GD82G. Section within Final Report: Forecasting Rangeland Condition with GIS in Southeastern Idaho.

Wen, Q., Ge, Z. & Song, Z. (2015). Multimode dynamic process monitoring based on mixture canonical variate analysis model. *Industrial and Engineering Chemical Research* **54**, 1605–1614.

Wu, G., Holan, S. H., Nilon, C. H. & Wikle, C. K. (2015). Bayesian binomial mixture models for estimating abundance in ecological monitoring studies. *The Annals of Applied Statistics* **9**, 1–26.

Xie, X. & Shi, H. (2012). Dynamic multimode process modeling and monitoring using adptive Gaussian mixture models. *Industrial and Engineering Chemical Research* **51**, 5497–5505.

Yu, J. (2012a). A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. *Chemical Engineering Science* **68**, 506–519.

Yu, J. (2012b). A particle filter driven dynamic Gaussian mixture model approach for complex process monitoring and fault diagnosis. *Journal of Process Control* **22**, 778–788.

Yu, J. & Quin, S. J. (2008). Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE Journal* **54**, 1811–1829.

Yu, J. & Quin, S. J. (2009). Multiway Gaussian mixture model based multiphase batch process monitoring. *Industrial and Engineering Chemical Research* **48**, 8585–8594.