

## **Semiparametric regression for count data**

BY CINZIA CAROTA

*Department of Statistics and Applied Mathematics 'D. De Castro', Università di Torino,  
Piazza Arbarello, 8, 10122 Torino, Italy*  
carota@cisi.unito.it

AND GIOVANNI PARMIGIANI

*Department of Oncology, Johns Hopkins University, 550 North Broadway, Suite 1103,  
Baltimore, Maryland 21205, U.S.A.*  
gp@jhu.edu

### SUMMARY

We introduce a class of Bayesian semiparametric models for regression problems in which the response variable is a count. Our goal is to provide a flexible, easy-to-implement and robust extension of generalised linear models, for datasets of moderate or large size. Our approach is based on modelling the distribution of the response variable using a Dirichlet process, whose mean distribution function is itself random and is given a parametric form, such as a generalised linear model. The effects of the explanatory variables on the response are modelled via both the parameters of the mean distribution function of the Dirichlet process and the total mass parameter. We discuss modelling options and relationships with other approaches. We derive in closed form the marginal posterior distribution of the regression coefficients and discuss its use in inference and computing. We illustrate the benefits of our approach with a prognostic model for early breast cancer patients.

*Some key words:* Generalised linear model; Marginal model; Product of Dirichlet process mixtures.

### 1. INTRODUCTION

We consider a Bayesian semiparametric approach for modelling regression problems in which the response variable is defined on a countable set and the predictors are either categorical or quantitative with a finite number of possible values.

Our motivating application, discussed in detail in §4, concerns the prediction of the number of metastases of the axillary lymph nodes in patients with early-stage breast cancer. For each combination of predictors, or covariate profile, or cell, we are interested in modelling the distribution of the number of lymph node metastases found during surgery. There is biological support for the notion that the number of metastases should increase with the size of the primary tumour, and should decrease with age. This motivates smoothness in the prognostic variables' effects. Within each cell, that is conditional on covariates, typical empirical distributions of the response present heavy tails and a substantial number of zeros, corresponding to patients with no detectable metastatic disease. Zeros are more numerous than would be predicted by a Poisson model, and the degree

of this zero inflation depends on the covariates. While some cells are very well represented in the sample, others only contain a few observations. Therefore, a prognostic model based solely on the cell's empirical distributions would be unreliable in many of the cells, despite the large size of the overall sample. On the other hand, a parametric model would be too restrictive in well-populated cells.

Regression for count data is traditionally approached using generalised linear models (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989). Here we extend generalised linear models to accommodate arbitrary forms of the response distribution, and thus handle in a convenient, general and robust framework common problems such as overdispersion, underdispersion, zero inflation and heteroscedasticity. We develop our extension from a Bayesian viewpoint, to provide an accurate assessment of parameter and prediction uncertainty, without relying on asymptotic approximations, and to produce inferences in the form of probability distributions, for use in subsequent decision analyses.

We represent the relationship between covariates and response by a parametric relationship, while allowing for flexible forms for the response distribution. Flexibility refers to both the shape of the distribution and to how this shape changes as a function of the covariates. To achieve it, we assume that the cumulative distribution function of the response in each cell is unknown, and is described by a Dirichlet process mixture (Antoniak, 1974). Further progress in Dirichlet process modelling and computations is reviewed by Ferguson et al. (1992), Dey et al. (1998), MacEachern (1998) and Escobar & West (1998).

Since the Dirichlet process assigns all its mass to distributions defined on countable sets, it is especially attractive for analysing discrete quantitative data in multivariate settings. Each Dirichlet process can be characterised by a mean cumulative distribution function and a scalar dispersion parameter, controlling the amount of variation of the unknown cumulative distribution function around its mean. In a Dirichlet process mixture, both the mean cumulative distribution function and the dispersion parameters are considered unknown. Here, we model the mean cumulative distribution function using a parametric form, whose parameters are in turn specified to be parametric functions of the covariates. We also model the dispersion parameter as a parametric function of the covariates. In discrete-data models, observations are informative about the dispersion parameter of a Dirichlet process. As a result, the degree to which predictions and parameter estimates deviate from those obtained under the mean parametric model depends on the observations, via learning about the dispersion parameter. In the breast cancer application, our model will uncover structure and model sparse cells as in a Poisson regression, while providing predictive distributions that are close to the empirical distribution in cells with large numbers of observations.

Modelling of the mean of a Dirichlet process mixture as a function of predictors has precedents in Cifarelli (1979) and in an Istituto Matematico 'G. Castelnuovo', Rome, technical report by D. M. Cifarelli, P. Muliere and M. Scarsini; see also Poli (1985), Carota (1988) and Muliere & Petrone (1993) among others. Within the Bayesian approach, successful extensions of generalised linear models have been developed using hierarchical random effect models in which the distribution of the random effects can be described using parametric forms (Wong & Mason, 1985; West, 1985; Zeger & Karim, 1991; Albert & Chib, 1993) or more flexible semiparametric specifications (Bush & MacEachern, 1996; Müller & Rosner, 1997; Mukhopadhyay & Gelfand, 1997; Ibrahim & Kleinman, 1998; Escobar & West, 1998). In all of these approaches it can be challenging to model distributions of random effects, mixture components or other latent variables.

In the presence of multiple sources of heterogeneity, as in our breast cancer application, one has either to tackle a complex modelling effort at the second stage, or run the risk of significant model misspecification. Also, adequately flexible modelling (Müller & Rosner, 1997; Mukhopadhyay & Gelfand, 1997) can involve a substantial computational burden. The family of models proposed here is complementary to these approaches, as the nonparametric specification is made directly on the distribution of the observations, and does not require the specification of a random effect distribution.

Our approach is general in specifying the response distribution, as it assigns prior mass to all possible marginal models, and thus includes as special cases the marginal distributions resulting from random effect models. This flexibility is coupled with relatively simple analytical requirements. A result discussed in § 3 provides a closed-form expression for the marginal posterior distribution of the parameters of interest. This enables efficient sampling from the marginal posterior distribution of the covariates' coefficients, bypassing the need to sample random effects, mixture components or other latent variables. This simplifies prediction and inference about the unknown cumulative distribution functions of the response. Prior specification is low-dimensional and relatively simple in this context; reasonable default choices are available. No modelling commitment is necessary, other than that on the traditionally interpretable parametric backbone.

## 2. A SEMIPARAMETRIC MODEL FOR COUNT DATA

### 2.1. Notation and model assumptions

Consider a dataset that records, for each of  $N$  units, a response variable with countable support  $\mathcal{Y}$  and a vector of  $p - 1$  predictors that are either categorical or quantitative with a finite set of possible values. We indicate by  $X$  the  $N \times p$  design matrix of the observed values of the predictors and a column for the intercept, by  $x'_i$  the row corresponding to the  $i$ th unit, with  $i = 1, \dots, N$ , and by  $y$  the  $N$ -dimensional vector of values of the response variable. A predictor profile, or cell, is a combination of the observed values of the predictors. Let  $k$  index the profiles, and let  $K \leq N$  be the number of profiles. Let  $k(i)$  be the profile corresponding to unit  $i$ .

We assume that all units in a cell are exchangeable, with unknown distribution  $F_k$ . Modelling of the  $F_k$ 's has two goals: to represent the relationship between covariates and response via relatively simple parametric relationships; and to specify a flexible model for the shape of the  $F_k$ 's with regard both to deviations from parametric forms and to cell-to-cell variations. The first stage is to model each  $F_k$ , conditional on parameters, as a separate Dirichlet process  $\mathcal{D}(A_k, F_{0k})$  (Ferguson, 1973), parameterised by its mean function  $F_{0k}$  and total mass parameter  $A_k$ .

The cell-specific distributions are related through their mean cumulative distribution functions in the following way. Let  $F_0(\cdot|\theta)$  be a parametric distribution, such as the binomial, the beta binomial, the hypergeometric or the Poisson distribution, with  $\theta$  unknown. Each of  $\theta$ 's  $D$  dimensions will be modelled as an invertible function  $g_d$  of linear combinations of the predictors. In this way, the cumulative distribution function  $F_k$  of the response in each cell is distributed, conditional on  $\theta$ , according to a Dirichlet process with cell-specific mean function  $F_{0k} = F_0(\cdot|\theta_k)$ , with  $\theta_k = (\theta_{1k}, \dots, \theta_{Dk})$ . The corresponding probability distribution function will be indicated by  $f_{0k} = f_0(\cdot|\theta_k)$ . We refer to the  $F_{0k}$ 's as the parametric backbone of our model. The total mass parameter  $A_k$ , controlling the amount of variation of each of the cell-specific cumulative distribution functions around the parametric backbone, is also modelled as a function of a linear combination of the predictors.

In summary, the class of models that we have just defined can be represented as follows:

$$y|F \sim \prod_{i=1}^N F_{k(i)}(y_i) \quad (F_k \in \mathcal{F}), \quad (1)$$

$$F|\theta_1, \dots, \theta_K, A_1, \dots, A_K \sim \prod_{k=1}^K \mathcal{D}\{A_k, F_0(\cdot|\theta_k)\},$$

$$g_0(A_k) = x'_k \gamma \quad (k = 1, \dots, K), \quad g_d(\theta_{ak}) = x'_k \beta_d \quad (k = 1, \dots, K; d = 1, \dots, D),$$

$$\beta, \gamma \sim \pi(\beta, \gamma).$$

In the sequel, model (1) will indicate the entire model specification above.

Here  $\mathcal{F}$  is the set of all cumulative distribution functions on  $\mathcal{Y}$  and  $F = (F_1, \dots, F_K)$ . For each cell we have an unknown cumulative distribution function  $F_k$  and  $D + 1$  parameters  $(\theta_{1k}, \dots, \theta_{Dk}, A_k)$  indexing the conditional Dirichlet process of  $F_k$ . The  $(D + 1)k$  cell-specific parameters are functions of the  $(D + 1)p$  coefficients  $\beta$  and  $\gamma$ . The  $F_k$ 's are assumed to be conditionally independent given  $\beta$  and  $\gamma$ . In many applications, the focus will be on posterior inference for the  $(D + 1)p$  coefficients and on prediction of the response for a future unit in a specified cell, in which case  $F$  can be considered as a nuisance parameter. Since  $\beta$  and  $\gamma$  are unknown and are assigned a prior distribution, the resulting overall specification is a product of mixtures of Dirichlet processes (Cifarelli & Regazzini, 1978).

As we will discuss in more detail in § 3, within a cell the marginal likelihood function for  $A_k$  may not be integrable as a function of  $A_k$ , and can be very skewed. The reason is that large values of  $A_k$  correspond to distributions that are very close to the parametric model, and it is difficult to distinguish these from each other if the sample size in the cell is not large. We suggest using parsimonious parameterisations for the  $A_k$ 's that can be estimated based on a relatively small number of well-populated cells.

## 2.2. Examples

*Example 1.* In the lymph node metastases example, outlined earlier, we implement a Bayesian semiparametric version of Poisson regression. The response  $y$  is the number of metastases found, so  $\mathcal{Y}$  is the set of positive integers,  $F_0$  is the Poisson family, so that  $D = 1$ , and  $\theta$  is the mean of the Poisson distribution. The function  $g_1$  plays the role of the link function in generalised linear models; in our example  $g_1 = \log$ . A cell corresponds to a specific profile of covariates, one example being women whose primary cancer was diagnosed between ages 50 and 54 years, was 2 to 3 millimetres in size and was early-stage. The parameter  $A_k$  controls the degree to which the cell-specific  $F_k$  departs from that of the Poisson. We also take  $g_0 = \log$ . This specification can accommodate a broad range of departures from Poisson regression.

*Example 2.* Overdispersed generalised linear models can also be seen as special cases of the strategy described here. An example is the beta binomial model, for which

$$f_0(y|\alpha, \beta) = \binom{N}{y} \frac{B(y + \alpha, N - y + \beta)}{B(\alpha, \beta)} \quad (y = 0, \dots, N).$$

A useful reparameterisation is  $\theta_1 = \alpha/(\alpha + \beta)$  and  $\theta_2^2 = 1/(\alpha + \beta + 1)$ . The mean and variance are given by  $E(Y) = N\theta_1$  and

$$\text{var}(Y) = N\theta_1(1 - \theta_1)\{1 + (N - 1)\theta_2^2\}.$$

The factor  $\{1 + (N - 1)\theta_2^2\}$  controls additional dispersion compared to the binomial model.

Liang & McCullagh (1993) highlight the fact that inference about the parameters of interest can be sensitive to alternative models for overdispersion, of which the beta binomial model is one. In this regard, a semiparametric specification such as that proposed here is attractive, because it allows for arbitrary patterns of under- and over-dispersion, and bypasses delicate model selection, which requires artful diagnostics and pre-testing.

Beta binomial backbones allow one to model the relationship between overdispersion and covariates while maintaining a flexible overall specification. For example, an attractive parameterisation for the dispersion parameter is to set  $g_2(\theta_{2k}) = \log\{1 + (N - 1)\theta_{2k}^2\}$ . In model (1), increased dispersion compared to the binomial model is also introduced by the total mass parameter  $A$ . While  $\theta_2$  accounts for dispersion of the kind generated by a beta mixture,  $A$  can also control additional features, such as excess zeros or bimodalities. Detecting differences between these patterns of overdispersion, by estimating both  $A_k$  and  $\theta_{2k}$  as a function of covariates, requires large sample sizes. Since model (1) can accommodate over- and under-dispersion whether the mean cumulative distribution function  $F_0$  is binomial or beta binomial, the additional flexibility afforded by a beta binomial  $F_0$  is important mainly when the parameter  $\theta_2$  and its relation to covariates are of direct scientific interest.

*Example 3.* Robust extensions of Bayesian logistic regression can be developed within the family (1). Consider a simple example which is common in developmental toxicology applications (Ryan, 1992). Each of  $N$  dams is implanted with  $n$  offspring and receives a dose  $d_k$  of a chemical being tested for toxicity. The response is the count  $y_i$  ( $i = 1, \dots, N$ ) of the number of offspring that present a characteristic of interest. An ordinary logistic regression with exchangeable offspring would postulate  $y_i \sim \text{Bi}(\theta_{k(i)}, n)$ , with  $\text{logit}(\theta_{k(i)}) = \beta_0 + \beta_1 d_{k(i)}$ . A more robust Dirichlet process mixture elaboration can be formed by setting  $y_i \sim F_k$  with  $F_k \sim \mathcal{D}\{\text{Bi}(\theta_k, n), A_k\}$ , and with  $\theta_k$  modelled as above (Dominici & Parmigiani, 2001).

An alternative Bayesian hierarchical model uses random effects for the dams. A version of the latter is  $y_i \sim \text{Bi}(\theta_i, n)$ , with  $\text{logit}(\theta_i) = \beta_{0i} + \beta_1 d_{k(i)}$ . In turn, variability in the dam-specific intercepts  $\beta_{0i}$  can be modelled by a parametric form, or more flexibly by a Dirichlet process mixture (Mukhopadhyay & Gelfand, 1997). In model (1), a Dirichlet process mixture is assigned to the unknown cumulative distribution functions in each of the  $K$  dose levels. The mixing distributions of the Dirichlet process mixtures are then connected at a higher level via the dose effect. In the random-effect approach, a Dirichlet process mixture is assigned to the unknown distribution of the random effect. Our approach is simpler computationally, and potentially more flexible in modelling the distribution of the response within each dose level. On the other hand it does not address the problem of estimating the dam-to-dam variability, which may be of scientific interest in some applications.

*Example 4.* Developmental toxicology applications sometimes require joint consideration of multiple outcomes in the offspring, if for example there are two types of birth defect (Lefkopoupou et al., 1989). The response is then a pair of counts  $(y_{1i}, y_{2i})$ , with  $i = 1, \dots, N$ , of the number of offspring that present a characteristic of interest. Since the distribution of the response is finitely supported, the Dirichlet process assumption is equivalent to a Dirichlet distribution. Given any parametric backbone, such as the polytomous regression model, we can specify a model in our class by assuming that, for each

covariate profile, the parameters  $p(y_{1i}=j_1, y_{2i}=j_2)$ , for  $0 \leq j_1 \leq N$  and  $0 \leq j_2 \leq N$ , follow a Dirichlet distribution with mean defined by the parametric backbone.

*Example 5.* Leonard & Novick (1986) consider a flexible Bayesian formulation for modelling cell probabilities in a two-by-two table. They assume that cell counts are Poisson with parameter  $\varphi_{yk}$ , where the pair  $(y, k)$  indexes a cell in the table. At the second stage of their model they specify  $\varphi_{yk}$  to be independent gamma random variables with mean  $\theta_{yk}$  and variance  $\theta_{yk}/A$ . The similarities between their model and ours are highlighted by our notation that suggests thinking of one of the dimensions of the table, the rows say, as the response, and the other as the covariate profile. An important difference is that the cell counts, which in our case correspond to the number of replicate observations in each covariate profile, are Poisson in Leonard & Novick (1986) and are generated by a more general urn-type model in our specification 1; see Antoniak (1974) for details.

*Example 6.* Models of the form (1) can be used for discrete survival analysis. A constant hazard rate parametric backbone is provided by the geometric distribution. The log of the hazard rate can be modelled as a function of covariates. Censored data add complexity, as the useful closed-form marginalisation of the likelihood breaks down. Computations are still feasible via data augmentation, as shown by Giudici et al. (2002) for a continuous response.

### 3. DISTRIBUTIONAL RESULTS

#### 3.1. Marginal posterior distribution

A key feature of our modelling strategy is that we can write a closed-form expression for the marginal posterior probability distribution of the parameters  $\beta$  and  $\gamma$ , by integrating out  $F$ . We use  $\tilde{X}$  to denote the matrix of unique rows of  $X$ ; the generic row is a profile of predictors, denoted by  $\tilde{x}_k$ , for  $k = 1, \dots, K \leq N$ . The number of observations with profile  $k$  is  $N_k$ . We also use  $\tilde{y}_k$  to denote the vector of  $r_k$  unique observations in  $y$  for predictor profile  $k$ ;  $N_{jk}$  counts number of occurrences of  $\tilde{y}_{jk}$ . Also,  $x_{(n)} = x(x+1) \dots (x+n-1)$  and  $x_{(0)} = 1$ .

Extending the technique described in Lemma 1 of Antoniak (1974) to our specification, see also Cifarelli & Regazzini (1978), we can integrate out the unknown  $F_k$ 's from the sampling distribution, and thus derive the marginal likelihood function of  $(\beta, \gamma)$ . The extension entails one application of Lemma 1 for each covariate profile, and is made possible by the conditional independence of the profiles given  $\beta$  and  $\gamma$ . The resulting marginal likelihood is

$$L(\beta, \gamma) \propto \prod_{k=1}^K (A_k)_{(N_k)}^{-1} \prod_{j=1}^{r_k} A_k f_0(\tilde{y}_{jk} | \theta_k) \{A_k f_0(\tilde{y}_{jk} | \theta_k) + 1\}_{(N_{jk}-1)}, \quad (2)$$

where  $A_k = g_0^{-1}(x'_k \gamma)$  and  $\theta_{dk} = g_d^{-1}(x'_k \beta_d)$ . The joint posterior distribution of  $F$  and  $(\beta, \gamma)$  can then be factorised analytically into the marginal posterior distribution of  $(\beta, \gamma)$  and the posterior distribution of  $F$  given  $(\beta, \gamma)$ :

$$\pi(\beta, \gamma, F | y) = \pi(\beta, \gamma | y) \pi(F | \beta, \gamma, y) \propto \pi(\beta, \gamma) L(\beta, \gamma) \pi(F | \beta, \gamma, y). \quad (3)$$

Also, for each predictor profile, and given  $\beta$  and  $\gamma$ , the  $F_k$ 's are a posteriori conditionally independent and follow a Dirichlet process with total mass parameter  $A_k + N_k$  and mean parameter

$$\tilde{F}_k = \zeta_k F_0(\cdot | \theta_k) + (1 - \zeta_k) \hat{F}_k,$$

where  $\hat{F}_k$  is the empirical cumulative distribution function in cell  $k$ , and the weight

$$\zeta_k = \frac{A_k}{A_k + N_k}$$

controls the amount of shrinkage that will be applied to the posterior means of the cell-specific cumulative distribution functions in cell  $k$ . The joint posterior distribution can be rewritten as

$$\pi(\beta, \gamma, F | y) \propto \pi(\beta, \gamma) L(\beta, \gamma) \prod_{k=1}^K \mathcal{D}(A_k + N_k, \tilde{F}_k). \quad (4)$$

If we use this factorisation, it is simple to focus on marginal inference on  $\beta$  and  $\gamma$ , by sampling from  $\pi(\beta, \gamma | y)$ , or to generate from the posterior and predictive distribution at the cell level.

### 3.2. Relationships with parametric models

The likelihood function corresponding to the parametric backbone is

$$L_0(\beta) = \prod_{k=1}^K \prod_{j=1}^{r_k} p_0(\tilde{y}_{jk} | \theta_k)^{N_{jk}}. \quad (5)$$

*Remark 1.* If, for all  $k$ ,  $N_k = 1$ , then

$$L(\beta, \gamma) = L_0(\beta).$$

When each subject has a separate predictor profile, the marginal likelihood of the semi-parametric model is independent of the total mass parameters  $A_k$ , and therefore there is no learning about the parameter  $\gamma$ . This highlights the fact that models of the form (1) are not well suited to applications in which the predictors are not discrete or cannot be coarsely discretised without loss of information.

*Remark 2.* If, for all  $j$  and  $k$ ,  $N_{jk} = 1$ , then

$$L(\beta, \gamma) = L_0(\beta) L_1(\gamma),$$

where

$$L_1(\gamma) = \prod_{k=1}^K \frac{A_k^{N_k}}{(A_k)_{(N_k)}}.$$

If there are no duplicate observations in any of the cells, inference on  $\beta$  is the same as it would be under the parametric model. In this case the information is sparse and it is not useful to model separate cumulative distribution functions in each cell. In both this case and the one above there are not enough observations to provide information about the cell-specific cumulative distribution functions, and model (1) reverts to the parametric case.

*Remark 3.* If, for all  $k$ ,  $A_k \rightarrow \infty$ , then

$$L(\beta, \gamma) \rightarrow L_0(\beta).$$

With larger values of the  $A_k$ 's the  $F_k$ 's tend to be closer to their mean,  $F_{0k}$ . The parametric backbone is therefore a limiting case of model (1), obtaining when the prior on  $\gamma$  assigns mass to large values of the  $A_k$ 's. Generalised linear models also obtain as a limiting case of (1) when in addition  $F_0$  belongs to an exponential family. Furthermore, for fixed  $\beta$ , the

tail of the marginal likelihood  $L$  converges to a horizontal asymptote as the  $A_k$ 's become large. This asymptote, while often very small in applications, is not exactly zero. The data do not provide information to discriminate among very large values of the  $A_k$ 's. The posterior distribution on  $\gamma$  will not necessarily be proper unless the prior is proper.

### 3.3. Priors

Our approach only requires the specification of a prior distribution on  $\beta$  and  $\gamma$ . With respect to  $\beta$ , dispersed but proper priors accompanied by sensitivity analysis will work in many applications when the focus is on estimation and prediction rather than testing. If it is easy to interpret the components of  $\theta$ , elicitation of the prior on  $\beta$  can be guided by knowledge of the likely magnitude of the predictors' effects, as in conventional generalised linear models.

The choice of prior distribution on  $\gamma$  can be more delicate. The  $A_k$ 's control the amount of deviation from the mean parametric model. The more mass is assigned to values of  $\gamma$  leading to large values of the  $A_k$ 's, the more the model will be close to its parametric backbone. The prior on  $\gamma$  can therefore be used to specify the degree of flexibility that is desired of the response distributions. In general, even though the data can provide relevant information about  $\gamma$ , it is important to specify a proper prior on  $\gamma$ , in view of Remark 3 in § 3.2.

A practical strategy for assigning a prior distribution on  $\gamma$  is to consider the implied distribution on the shrinkage weights  $\zeta_k$ . In the context of Example 5, Leonard & Novick (1986) consider the amount of shrinkage that will be applied to the posterior means of the cell probabilities. In their model, this does not depend on the covariate profile  $k$ , and there are potentially useful default prior specifications. For example, they suggest using a uniform distribution on their shrinkage parameter  $1/(1 + A)$ .

This is an attractive strategy, but to use it in our case would correspond to specifying the prior in terms of the shrinkage expected in a cell with a single observation. A simple extension is to specify the prior on  $\gamma$  by assuming, first, a dispersed prior with mean zero on the coefficients  $\gamma_1, \dots, \gamma_D$ , and, secondly, a uniform distribution on  $N^*/(A + N^*)$ , where  $N^*$  is the sample size in a 'target' cell;  $N^*$  could be a plausible sample size, or one for which it is natural or desirable to assume a uniform distribution on the unknown shrinkage weight. The resulting distribution of  $\zeta_k$  has density  $N^*N_k/\{\zeta_k N^* + (1 - \zeta_k)N_k\}$ . When  $g_0 = \log$ , conditional on  $\gamma_1 = 0, \dots, \gamma_D = 0$ , the uniform distribution on  $N^*/(A + N^*)$  translates to a logistic distribution for  $\gamma_0$ ; that is

$$p(\gamma_0 | \gamma_1 = 0, \dots, \gamma_D = 0) = N^*e^{\gamma_0}/(N^* + e^{\gamma_0}).$$

The location parameter is  $\log(N^*)$  while the scale parameter is one. This is usually a reasonable compromise, that will penalise cases that are close to the fully parametric and fully nonparametric cases, but will allow for adequate flexibility in the intermediate range. An alternative strategy, closely related in intention, is illustrated in § 4.

When the prior on  $\gamma$  gives mass to values of  $\zeta$ 's near to 1, our model can be interpreted as a small-neighbourhood elaboration, or robustification, of the parametric model. This can be used to construct an all-purpose diagnostic of sensitivity to the choice of parametric model. Starting with a currently entertained parametric model, we implement our small-neighbourhood elaboration and reassess estimates of interest. When large deviations in the conclusions occur even for priors on  $\zeta_k$  that are highly concentrated around 1, this suggests the need for a more general model (Carota et al., 1996).

#### 4. PROGNOSTIC MODELLING OF AXILLARY NODE INVOLVEMENT IN BREAST CANCER

Axillary lymph node dissection is a common additional surgical procedure for early breast cancer patients receiving breast conservation surgery. Its value is in the prognostic information and as an aid in choosing an appropriate adjuvant therapy after surgery (Recht & Houlihan, 1995). It is potentially important to predict accurately axillary lymph node status of information available at the time of the initial diagnosis, because many patients could avoid lymph-node surgery and the resulting morbidity (Ravdin et al., 1994; Parmigiani et al., 1999).

Here we use a model from family (1) to develop a prognostic model for nodal status based on age, size of the primary tumour and histological stage. We consider data from the Surveillance, Epidemiology and End Results Program (National Cancer Institute, 1997). This registry is population-based and breast cancer cases are collected prospectively in several regions across the U.S.A. For each of  $N = 58\,695$  cases we consider the following variables: the age  $x_1$  of the patient at diagnosis, in 10-year intervals; the size  $x_2$  of the primary tumour, determined based on a diagnostic mammography before the axillary lymph node dissection and measured in 14 intervals, roughly uniform in the logarithmic scale; the histological stage  $x_3$  of the primary tumour, classified as a binary variable given by 'early stage', defined as stage I or II, or 'late stage' defined as stage III or IV; and the number  $y$  of surgically sampled axillary nodes that are positive after biopsy. A profile is a unique combination of  $x_1$ ,  $x_2$  and  $x_3$ . Our classification generates a total of 170 observed profiles.

The choice between a parametric and a semiparametric model involves a trade-off between the risk of overfitting the data and the risk of misspecifying the model. A practical approach for assessing model performance in this setting is out-of-sample validation, which mimics the use of the prediction model in actual clinical practice. Before our analysis we divided the cases into a 92% training sample and four 2% validation samples. The median number of patients per profile in the training sample is about 37, while the range is from 1 to 4363.

The Poisson distribution for the number of nodes within a profile would be appropriate if metastases in the lymph node occurred independently and at a constant rate. Both of these assumptions are unlikely to hold from a biological standpoint. Empirically, there are too many zeros and there are heavy tails. In terms of the relationship of the response to the covariates, the patterns of variability reflect a lower variance than predicted by the Poisson at small tumour sizes and higher at large tumour sizes.

For our analysis, we choose  $F_0$  to be a Poisson cumulative distribution function with mean  $\theta$ , and we choose a logarithmic link for both the mean response  $\theta_k$  and the total mass

$$y|F \sim \prod_{i=1}^N F_{k(i)}(y_i) \quad (F_{k(j)} \in \mathcal{F}), \quad (6)$$

$$F_k|\theta_k, A_k \sim \mathcal{D}\{A_k, \text{Po}(\cdot|\theta_k)\} \quad (k = 1, \dots, K),$$

$$\log(\theta_k) = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k}, \quad \log(A_k) = \gamma_0 + \gamma_1 x_{1k} + \gamma_2 x_{2k} + \gamma_3 x_{3k}.$$

The logarithm of the event rate is a standard choice for the link function in a Poisson regression. The parameterisation  $\log(A_k)$  can be expressed as the logit of a shrinkage factor  $\zeta_0 = A_0/(A_0 + A_k)$ , where the log of the unknown  $A_0$  is being absorbed into the intercept

$\gamma_0$ . An alternative interpretation of our specification is therefore that the shrinkage weight in a ‘typical’ cell depends on the covariates via a logistic regression relationship.

The prior distribution on  $\beta$  is normal, with mean 0 and diagonal covariance matrix, with elements (100, 0.22, 20, 0.31). Apart from the intercept, values are obtained by dividing 10, corresponding to a large change in the response, by the square of the likely interquartile range of the corresponding covariate. This is to obtain a proper but highly dispersed prior in which the spread of the regression coefficients is comparable across covariates.

We consider two families of prior distributions for  $\gamma$ . The first is normal with mean vector  $(\log(20), 0, 0, 0)$  and diagonal covariance matrix, with elements (1, 0.22, 20, 0.31). The distribution of  $\gamma_0$  was chosen by trial and error, visually inspecting distributions of  $\zeta_k$ ’s chosen to represent a range of likely sample sizes for the cells, and selecting parameters that ensure adequate a priori coverage of the unit interval for these cells. The implied prior distribution on the  $A_k$ ’s is a log-normal. The median is a common value 2.99 for each covariate profile, because the prior means of  $\gamma_1, \gamma_2$  and  $\gamma_3$  are zero. The variances of the  $A_k$ ’s and  $\zeta_k$ ’s vary with  $k$ . To give an idea of the implied variability in the  $\zeta_k$  scale, we computed 98% probability intervals and standard deviations for all profiles. The lower bounds have a median of  $1.298 \times 10^{-10}$  and interquartile range  $1.534 \times 10^{-21}$  to  $2.064 \times 10^{-8}$ , while the upper bounds have a lower quartile greater than 0.99999. The standard deviation averaged 0.43. In summary, this specification does not appear to impose strong a priori information on the shrinkage weights  $\zeta_k$ . The second family of priors on  $\gamma$  replaces the above distribution on  $\gamma_0$  with the logistic distribution described in § 3.3, with  $N^* = 20$ . This prior can be thought of as a default choice, and is used as our baseline case in discussing results and sensitivity analyses.

We generated a sample from the joint distribution of  $\beta$  and  $\gamma$  using a Metropolis–Hastings algorithm (Tierney, 1994) with random walk proposal. Since the dimensionality of the parameter space is relatively low, we found it efficient to simulate joint moves for the whole vector  $(\beta, \gamma)$ . After initial exploration of the posterior, the proposal covariance matrix was estimated based on the empirical covariance matrix in early runs. Results presented here are based on an equally spaced subset of 1000 observations from a chain of 10 000. Standard convergence diagnostics do not reveal convergence problems.

We compared model (6) to an overdispersed generalised linear model with Poisson error in which the event rate in profile  $k$  is  $\theta_k^\phi$ , and, as in model (6),  $\log(\theta_k) = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k}$ . We used the same priors on  $\beta$  and a flat prior on the overdispersion parameter  $\phi$ , and obtained a Markov chain Monte Carlo sample. An intermediate level of flexibility between the two models could be achieved using mixtures of Poisson distributions (Viallefont et al., 2002). These would provide a practical approach for handling excess zeros, although it may be complicated to fit mixtures that adapt to the different covariate profiles as flexibly as model (6). Parsimonious mixtures of Poisson distributions could also be used as the parametric backbone to a model like (6). We would recommend this primarily if the mixture-specific parameters were of direct scientific interest.

Table 1 summarises the posterior means and standard deviations of the regression coefficients for both models. The two models make different assumptions about both the distribution of the response and the dependence of the response mean and variance on covariates. Coefficients play different roles and have different interpretations, and differences are not necessarily attributable to bias. The most pronounced differences are in the effect of the stage variable, which is captured by the rate in the generalised linear model but only affects the total mass parameter in model (6). We also carried out simulations,

not shown here in detail, using data generated from a Poisson regression model, and obtained a very close agreement between the regression coefficients in the standard and semiparametric models.

Table 1. *Posterior means and standard deviations of the parameters  $\beta$ ,  $\gamma$  and  $\phi$  in the product of Dirichlet process mixtures model (6), labelled PMDP, and in the overdispersed generalised linear model, labelled OGLM*

Parameter	PMDP model		OGLM model	
	Posterior mean	Posterior SD	Posterior mean	Posterior SD
$\beta_0$	2.38	0.00998	0.118	0.00569
$\beta_1$ (age)	-0.00217	0.000368	-0.0068	0.000277
$\beta_2$ (stage)	-0.588	0.0152	1.11	0.0301
$\beta_3$ (log tumour size)	0.0155	0.000717	0.0667	0.00177
$\gamma_0$	1.31	0.0403		
$\gamma_1$ (age)	-0.00301	0.00175		
$\gamma_2$ (stage)	1.35	0.0764		
$\gamma_3$ (log tumour size)	0.0291	0.00296		
$\phi$			1.02	0.0259

Figure 1 considers two covariate profiles, corresponding to a large,  $N_k = 3273$ , and a small,  $N_k = 5$ , sample size in the cell. For each, it shows a sample of realisations from the posterior distributions of the cell-specific probability distribution  $f_k$  and from the corresponding parametric estimate in the generalised linear model. The graphs illustrate the different degrees of adaptation of the nonparametric fit in the two cases. In Fig. 1(a) there is a large discrepancy between the semiparametric model and the overdispersed generalised linear model. The  $f_k$ 's display substantial zero-inflation, kurtosis and a heavy right tail. These features are missed by the generalised linear model's prediction. By contrast, in Fig. 1(b) the estimation of the cell-specific distribution borrows strength from other cells via the parametric backbone of model (6) and the associated regression relationship. The backbone is visible in the more dispersed U-shaped distribution, the only departures from which are the extra zeros and the higher mass assigned to  $y = 41$ , where an observation was made. In a small cell such as this the dispersion of predictions from model (6) will typically be greater than that of the generalised linear model's predictions. Figures 1(c), (d) show the distributions of the shrinkage weights  $\zeta_k$  in the two profiles. In the cell with a large sample the weight assigned to the parametric backbone is near zero, while in the cell with the small sample size it varies around 79%. The distributions of the weights  $\zeta_k$  are data-driven, and priors do not vary appreciably in the relevant range of the posterior. With smaller overall sample sizes this may not hold.

Table 2 summarises our out-of-sample validation. Our interest is in global measures of out-of-sample fit, in terms of both point prediction and predictive distributions. In addition, patient-level decision making may depend on specific features of the predictive distribution. For example, decisions regarding adjuvant therapy may depend on the probability of no nodal involvement ( $y = 0$ ); also, access to clinical trials of high-dose chemotherapy with bone-marrow transplantation are often restricted to patients with 10 or more nodes involved. In view of this, we considered five criteria, presented in Table 2: the root mean square of Pearson's residuals; the root mean square of deviance residuals; the root mean squared error in the prediction of the binary outcome  $y = 0$ ; the root mean squared error in the prediction of the binary outcome  $y \geq 10$ ; and the average log predictive

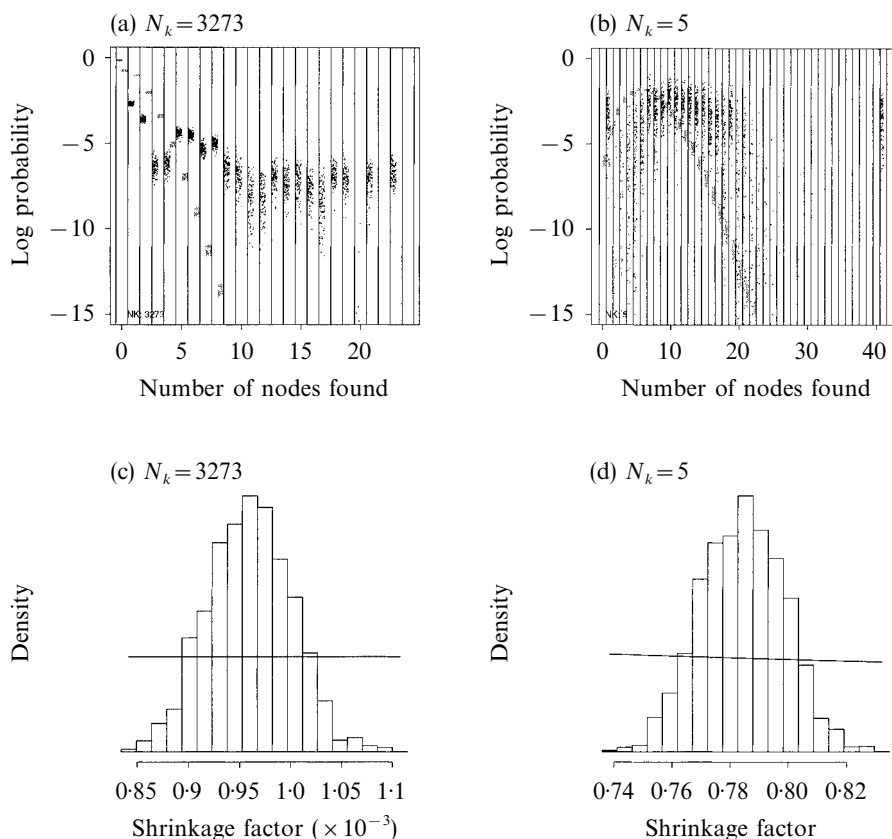


Fig. 1. Posterior predictive distribution and shrinkage factors for two covariate profiles. (a) and (b) display the posterior distributions of  $f_k$ : each vertical section corresponds to a value of  $y$  and contains two clouds of points, a sample from  $f_k(y)$  from model (6) (left) and a sample from the overdispersed generalised linear model (right and more lightly shaded). (c) and (d) display histograms of samples from the posterior distribution of the corresponding shrinkage factor  $\zeta_k$ , along with the prior density, renormalised to the range of the sampled values of the posterior. (a) and (c) refer to a profile with 3273 observations; (b) and (d) refer to a profile with 5 observations.

probability at the observed response, critical for assessing the accuracy of the whole predictive distribution (Bernardo & Smith, 1994, Ch. 5).

Model (6) has smaller Pearson's residuals and larger deviance residuals, with the gain in one type of residual roughly balancing the loss in the other. The superiority of the overdispersed generalised linear model in the deviance residual is to be expected, as it is in this case close to the generalised linear model, that is  $\phi$  is close to 1, and is thus close to minimising the deviance residuals. Model (6) provides better fit of the probability of  $y = 0$  and comparable fit on the right tail, taken overall. If we consider the whole predictive distribution, model (6) proves far superior, with improvements exceeding 30%. In conclusion, if interest is in point predictions the trade-off between model (6) and the overdispersed generalised linear model depends on the criterion chosen, while, if interest is in the entire predictive distribution, model (6) approach appears superior in this dataset.

Figure 2 provides a comparison of both deviance and Pearson's residuals for the two models in one of the validation samples. Model (6) produces systematically higher deviance residuals and systematically lower Pearson's residuals. On the other hand it lacks fit for

Table 2. Summary of out-of-sample validation results for the product of Dirichlet process mixtures model (6), labelled PMDP, and the overdispersed generalised linear model, labelled OGLM. Each row corresponds to one of the five following criteria: the root mean square of Pearson's residuals; the root mean square of deviance residuals; the root mean squared error in the prediction of the binary outcome  $y = 0$ ; the root mean squared error in the prediction of the binary outcome  $y \geq 10$ ; and the average log predictive probability at the observed response, labelled log. Each model is scored on four separate external validation sets

	OGLM model				PMDP model			
	Validation set				Validation set			
	1	2	3	4	1	2	3	4
Pearson	3.02	3.16	3.08	2.89	2.76	2.70	2.53	2.38
Deviance	2.00	2.05	2.12	2.05	2.25	2.25	2.30	2.23
pr( $y = 0$ )	0.56	0.56	0.56	0.54	0.51	0.49	0.50	0.52
pr( $y > 10$ )	0.23	0.22	0.25	0.22	0.22	0.24	0.25	0.24
log	-2.47	-2.60	-2.75	-2.65	-1.80	-1.79	-1.76	-1.93

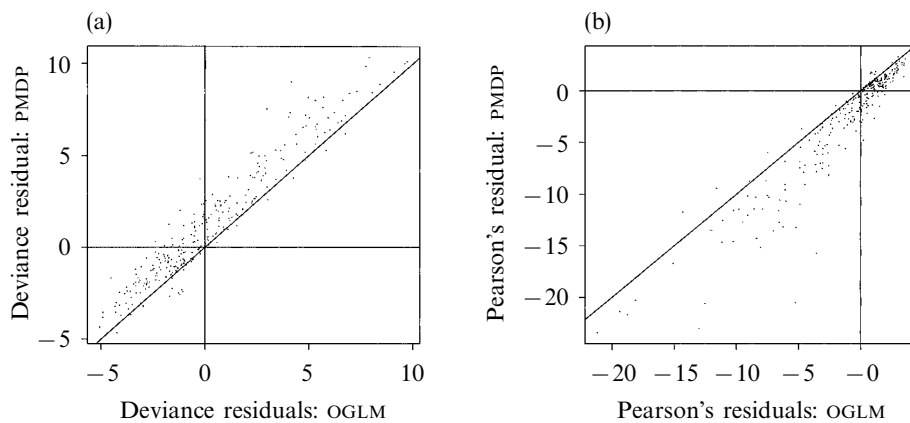


Fig. 2. Comparison of residuals. Deviance residuals (a) and Pearson's residuals (b) for the product of Dirichlet process mixtures, PMDP, model (6) (vertical scale) and the overdispersed generalised linear model, OGLM, (horizontal scale). Each point is one observation in the external validation sample number 3. Other validation samples produce similar results.

very small values, which determines some very large Pearson residuals. Figure 3 provides a comparison of the probabilities of the outcomes actually observed. Values above the diagonal indicate a better prediction from model (6). In a small number of cases at the top left, model (6) predicted the actual outcome with high probability, while the overdispersed generalised linear model gave it a very low probability. Most of these cases were zero, a critical outcome from a clinical standpoint. Model (6) performs better in an additional set of high-probability values above the diagonal, but is inferior in lower probability events.

Table 3 examines the sensitivity of the comparison of Table 2 to the choice of hyperparameters. Since we used dispersed priors for all parameters except possibly  $\gamma_0$ , we can focus the sensitivity analysis there. We first performed a sensitivity analysis by ten-fold changes in  $N^*$ , with virtually no change in the results. This is not reported here, as the distribution

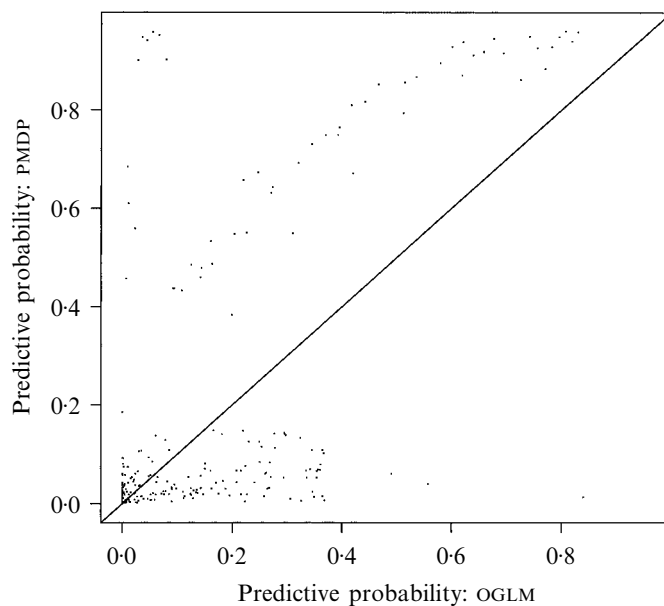


Fig. 3. Comparison of predictive probabilities. Predictive probabilities of observing the actual outcome in the validation set for the product of Dirichlet process mixtures, PMDP, model (6) (vertical scale) and the overdispersed generalised linear model, OGLM, (horizontal scale). Each point is one observation in the external validation sample number 3. Probabilities have been slightly perturbed to unmask repeated points.

Table 3. *Summary of sensitivity analysis on external validation results. Each row corresponds to one of the five following criteria: the root mean square of Pearson's residuals; the root mean square of deviance residuals; the root mean squared error in the prediction of the binary outcome  $y = 0$ ; the root mean squared error in the prediction of the binary outcome  $y \geq 10$ ; and the average log predictive probability at the observed response, labelled log. Scores are averaged over the four validation sets. Normal distributions are represented in terms of means and standard deviations*

	Baseline logistic	Scenario A $N\{\log(20), 1\}$	Scenario B $N\{\log(10\,000), 1\}$	Scenario C $N\{\log(20), 0.0001\}$	Scenario D $N\{\log(2000), 0.0001\}$
Pearson	2.59	2.59	2.59	2.53	2.63
Deviance	2.26	2.26	2.26	2.28	2.19
$\text{pr}(y = 0)$	0.50	0.50	0.50	0.51	0.51
$\text{pr}(y > 10)$	0.24	0.24	0.24	0.24	0.23
log	-1.82	-1.82	-1.82	-1.82	-1.92

of the shrinkage factors shows very little sensitivity to  $N^*$ . We then considered the normal case and examined four scenarios, namely baseline normal, Scenario A, small variance, Scenario B, large mean, Scenario C, and small variance and large mean, Scenario D. The only appreciable difference in performance is with Scenario D, which places high prior weight on a total mass parameter around 2000, and therefore brings the model closer to the parametric backbone and the validation results closer to those of the overdispersed generalised linear model. In this application, we have a very large sample and therefore

low sensitivity. In smaller samples we observe a sensitivity to the hyperparameters of the distribution of  $\gamma_0$ .

## 5. DISCUSSION

Our approach is motivated by several features: we can specify the distribution of the response in a general way, we can let the degree to which the distribution of the response adapts nonparametrically to the observations be determined by the data, we can obtain closed-form marginal likelihood and posterior distribution of the parameters of interest and use these to implement simple Markov chain Monte Carlo algorithms for prediction and inference; and we only need to elicit a prior distribution on a small number of parameters.

Our approach requires a parametric backbone, whose choice is an open question. We favour making the backbone complex when there is insight or interpretation to be gained from the extra parameters. Otherwise, the flexibility of family (1) should be sufficient to adapt to changes in the distribution. The dangers of misspecifying the backbone are likely to be less important than the dangers of misspecifying the model in a parametric analysis. Carota (1999) develops Bayes factors for weighting alternative backbones.

Quasilikelihood and generalised estimating equations procedures have provided simple alternatives for analysing count data (Williams, 1982; Liang & Zeger, 1986) without explicit commitment on the form of the response distribution, and without direct modelling of higher stage distributions of random effects. These often rely on asymptotic approximations. One goal of the methodology described here is to achieve simplicity and flexibility that may compete with those of generalised estimating equations approaches, while maintaining all the advantages of a full Bayesian analysis.

Like most semiparametric procedures, models from family (1) require moderate to large sample sizes in order to offer substantial advantages over parametric procedures. In general, they are likely to be worthwhile when there are multiple observations of the response in at least some of the predictor profiles. In particular, for inference on  $\beta$ , we recommend that at least  $pD$  cells should have an adequate number of observations for estimating the  $D$  elements of  $\theta_k$ . When data are limited, this model reverts the parametric backbone provided by the Dirichlet process mean.

## ACKNOWLEDGEMENT

This work was partly supported by the U.S.A.'s National Cancer Institute, within the Duke and Johns Hopkins Specialized Programs of Research Excellence in breast cancer, and by the Hecht scholar fund at Johns Hopkins Oncology Center. We thank Francesca Dominici, Tom Leonard and all reviewers for valuable comments.

## REFERENCES

- ALBERT, J. H. & CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assoc.* **88**, 669–79.
- ANTONIAC, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–74.
- BERNARDO, J. M. & SMITH, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.
- BUSH, C. A. & MACEachern, S. N. (1996). A semi-parametric Bayesian model for randomised blocked designs. *Biometrika* **83**, 175–85.

- CAROTA, C. (1988). Two-way layout: A nonparametric Bayesian approach (in Italian). In *Proceedings of the 34th Scientific Meetings of the Italian Statistical Society*, pp. 145–52. Siena: Nuova Immagine Editrice.
- CAROTA, C. (1999). Some results on Bayes factors in a nonparametric context. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 42–5. Alexandria, VA: American Statistical Association.
- CAROTA, C., PARMIGIANI, G. & POLSON, N. G. (1996). Diagnostic measures for model criticism. *J. Am. Statist. Assoc.* **91**, 753–62.
- CIFARELLI, D. M. (1979). Bayesian nonparametric approach of an analysis of variance problem (in Italian). *Annali dell' Istituto di Matematica Finanziaria dell'Università di Torino, Serie III* **17**, 1–20.
- CIFARELLI, D. M. & REGAZZINI, E. (1978). Nonparametric statistical problems under partial exchangeability. The use of associative means (in Italian). *Annali dell' Istituto di Matematica Finanziaria dell'Università di Torino, Serie III* **12**, 1–36.
- DEY, D., MÜLLER, P. & SINHA, D. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer.
- DOMINICI, F. & PARMIGIANI, G. (2001). Bayesian semi-parametric analysis of developmental toxicology data. *Biometrics* **57**, 166–73.
- ESCOBAR, M. & WEST, M. (1998). Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Ed. D. Dey, P. Müller and D. Sinha, pp. 1–22. New York: Springer.
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–30.
- FERGUSON, T. S., PHADIA, E. G. & TIWARI, R. C. (1992). Bayesian nonparametric inference. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, Ed. M. Ghosh and P. K. Pathak, pp. 127–50. Beachwood, CA: Institute of Mathematical Statistics.
- GIUDICI, P., MEZZETTI, M. & MULIERE, P. (2002). Mixtures of products of dirichlet processes for variable selection in survival analysis. *J. Statist. Plan. Infer.* To appear.
- IBRAHIM, J. G. & KLEINMAN, K. P. (1998). Semiparametric Bayesian methods for random effects models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Ed. D. Dey, P. Müller and D. Sinha, pp. 89–114. New York: Springer.
- LEFKOPOULOU, M., MOORE, D. & RYAN, L. (1989). The analysis of multiple correlated binary outcomes: Application to rodent teratology experiments. *J. Am. Statist. Assoc.* **84**, 810–5.
- LEONARD, T. & NOVICK, M. R. (1986). Bayesian full rank marginalization for two-way contingency tables. *J. Educ. Statist.* **11**, 33–56.
- LIANG, K.-Y. & MCCULLAGH, P. (1993). Case studies in binary dispersion. *Biometrics* **49**, 623–30.
- LIANG, K.-Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.
- MACEachern, S. N. (1998). Computational methods for mixture of Dirichlet process models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Ed. D. Dey, P. Müller and D. Sinha, pp. 23–44. New York: Springer.
- MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.
- MUKHOPADHYAY, S. & GELFAND, A. E. (1997). Dirichlet process mixed generalized linear models. *J. Am. Statist. Assoc.* **92**, 633–9.
- MULIERE, P. & PETRONE, S. (1993). A Bayesian predictive approach to sequential search for an optimal dose: Parametric and nonparametric models. *J. Ital. Statist. Soc.* **2**, 349–64.
- MÜLLER, P. & ROSNER, G. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *J. Am. Statist. Assoc.* **92**, 1279–92.
- NATIONAL CANCER INSTITUTE: SURVEILLANCE, EPIDEMIOLOGY, AND END RESULTS (SEER) PROGRAM (1997). SEER homepage. <http://www-seer.ims.nci.nih.gov>.
- NELDER, J. A. & WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc. A* **135**, 370–84.
- PARMIGIANI, G., BERRY, D. A., WINER, E. P., TEBALDI, C., IGLEHART, J. D. & PROSNITZ, L. (1999). Is axillary lymph node dissection indicated for early stage breast cancer — a decision analysis. *J. Clin. Oncol.* **17**, 1465–73.
- POLI, I. (1985). A Bayesian non-parametric estimate for multivariate regression. *J. Econometrics* **28**, 171–82.
- RAVDIN, P. M., DE LAURENTIIS, M., VENDELY, T. & CLARK, G. M. (1994). Prediction of axillary lymph node status in breast cancer patients by use of prognostic indicators. *J. Nat. Cancer Inst.* **86**, 1771–5.
- RECHT, A. & HOULIHAN, M. J. (1995). Axillary lymph nodes and breast cancer: A review. *Cancer* **76**, 1491–512.
- RYAN L. (1992). Quantitative risk assessment for developmental toxicity. *Biometrics* **48**, 163–74.
- TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with Discussion). *Ann. Statist.* **22**, 1701–62.
- VIALLEFONT, V., RICHARDSON, S. & GREEN, P. (2002). Bayesian analysis of Poisson mixtures. *J. Nonparam. Statist.* To appear.
- WEST, M. (1985). Generalized linear models: Scale parameters, outlier accommodation and prior distributions. In *Bayesian Statistics 2*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 329–48. Amsterdam: North-Holland.
- WILLIAMS, D. A. (1982). Extra-binomial variation in logistic linear models. *Appl. Statist.* **31**, 144–8.

- WONG, G. Y. & MASON, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *J. Am. Statist. Assoc.* **80**, 513–24.
- ZEGER, S. L. & KARIM, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *J. Am. Statist. Assoc.* **86**, 79–86.

[*Received March 2000. Revised July 2001*]