Chapter 2

Data Analysis

2.1. Density Estimation and Survival Analysis

The most straightforward application of BNP priors for statistical inference is in density estimation problems. Consider the generic density estimation problem, with data y_i, $i = 1, \ldots, n$, that is believed to be generated as an i.i.d. sample from some unknown distribution G. A BNP model can be used as a prior for G to complete the model

$$y_i \mid G \sim G \qquad\qquad G \sim p(G). \tag{2.1}$$

We could use, for example, a DP prior to specify $p(G)$ as $G \sim \mathsf{DP}(M, G_0)$, or a PT prior $G \sim \mathsf{PT}(\Pi, \mathcal{A})$. Many BNP models are conjugate under i.i.d. sampling. In other words, $p(G \mid y_1, \ldots, y_n)$ is in the same family as the prior, with updated parameters. This is true, for example, for the DP prior or the PT prior.

A limitation of many popular BNP models $p(G)$ for random probability measures is the discrete nature of G. This is the case, for example for the DP prior. A simple fix is the use of mixture models, convoluting the discrete RPM with a continuous kernel $f(x;\ \mu)$, e.g. $\mathsf{N}(x;\ \mu, 1)$,

$$y_i \mid F \sim F(y_i),\ F(y) = \int f(y;\ \theta)\ dG(\theta) \text{ and } G \sim p(G)$$

Such models are known as DP mixtures (DPM) etc. The model is illustrated in Figure 2.1. The point masses are the discrete probability measure G. Each point mass is smeared out with a kernel $f(x;\ \mu)$. The convolution of G and the kernel creates the continuous probability measure F. Posterior inference usually proceeds in an equivalent hierarchical model with latent variables $\theta_i \sim G$, $i = 1, \ldots, n$. The mixture is rewritten as

$$y_i \sim f(y_i;\ \theta_i) \qquad\qquad \theta_i \mid G \sim G \qquad\qquad G \sim p(G). \tag{2.2}$$

Posterior inference is still almost conjugate. If $p(G)$ was conjugate under i.i.d. sampling, then the complete conditional posterior $p(G \mid \theta_1, \ldots, \theta_n)$ for G given the imputed latent variables θ_i remains in the same family. And conditional on G the latent variables θ_i are usually easy to impute. We will discuss detail strategies in §3.3.

A special case of density estimation arises in survival analysis as density estimation with event time data, usually involving censoring. Survival analysis is a very traditional application of BNP in the early literature. Some BNP models for random probability measures remain conjugate even under (right) censoring. For example, a PT prior can be specified such that the posterior process for the unknown distribution remains a PT, even in the presence of censoring.

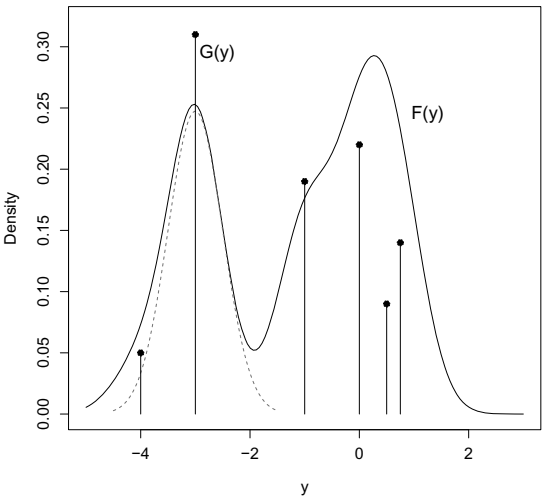


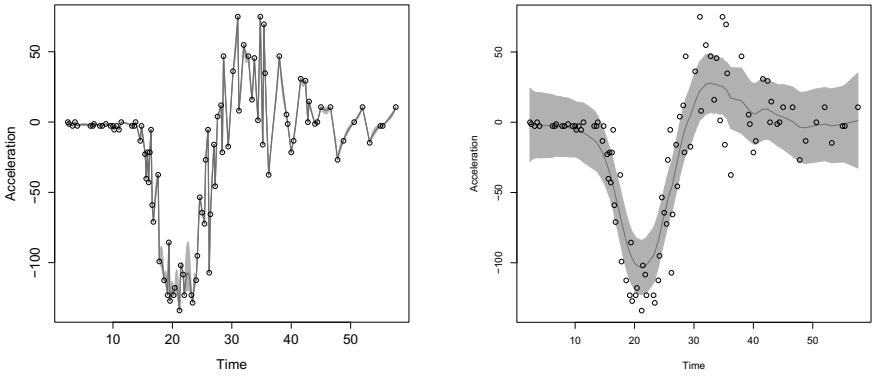
FIG 2.1. *Dirchlet Process mixture prior. The discrete random probability measure G is convoluted with a smooth kernel to create a continuous distribution F.*

2.2. Regression

Consider a generic regression problem with dependent variable y_i, covariates x_i, $i = 1, \ldots, n$, and an assumed model $y_i = f(x_i) + \epsilon_i$ with $\epsilon_i \sim p_\epsilon(\epsilon_i)$. As long as both, the regression function $f(\cdot)$ and the residual distribution $p_\epsilon(\cdot)$, are indexed by finitely many parameters, inference reduces to a traditional parametric regression problem. The problem becomes a non-parametric regression when the investigator wants to relax the parametric assumptions of either of the two model elements. This characterization of non-parametric regression allows for three cases.

2.2.1. Non-Parametric Residuals

The model can be generalized by going non-parametric on the residual distribution, assuming $\epsilon_i \sim G$ and a non-parametric prior $p(G)$, while keeping the regression mean function parametric as $f_\theta(\cdot)$ for a finite dimensional parameter vector θ. We refer to this case as a non-parametric error model. Essentially this becomes density estimation for the residual error. Of course the residuals ϵ_i are not usually observable. Hence, the problem reduces to one of density estimation conditional on assumed values for the parameters θ. A typical implementation using Markov chain Monte Carlo posterior simulation would include a transition probability that updates the currently imputed RPM G conditional on currently imputed values of θ. Conditional on θ the problem of updating inference on G reduces to density estimation for the residuals $\epsilon_i = y_i - f_\theta(x_i)$. And vice versa, conditional on imputing G, updating θ reduces to a regression problem with a known residual distribution.



(a) Interpolating Gaussian process (b) Smoothing Gaussian process

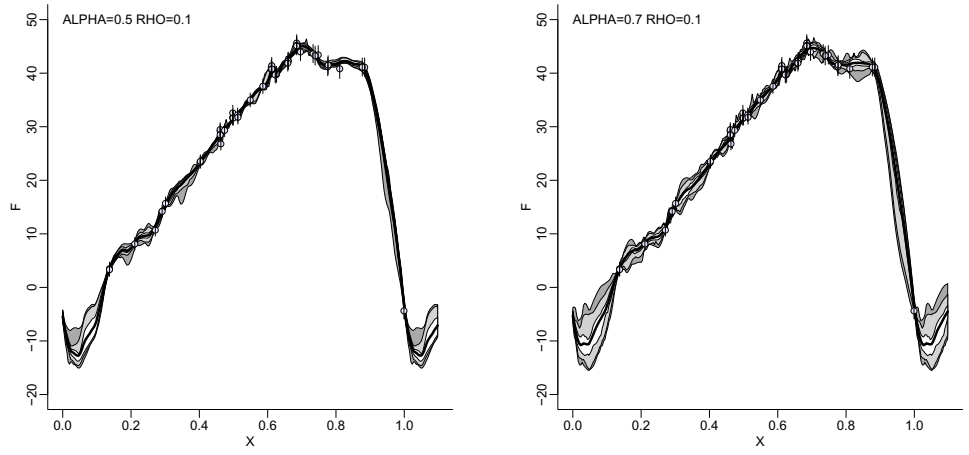
FIG 2.2. *An example of nonlinear regression using Gaussian processes. Points correspond to the observed data, the solid line corresponds to the posterior mean, and the grey bands (in panel (b)) are 95% credible intervals. The left panel corresponds to a model where* $\tau^2 = 0$*, so that the model acts as an interpolator. The model on the right panel allows for* $\tau^2 > 0$*, so that the predictor arising from the model behaves as a smoother.*

2.2.2. Non-Parametric Mean Function

Alternatively one could relax the parametric assumption on the mean function and complete the model with a non-parametric prior $f(\cdot) \sim p(f)$. We refer to this as a non-parametric regression mean function. As discussed in §1.3, popular choices for $p(f)$ are Gaussian process priors or priors based on basis expansions, such as wavelet based priors or neural network models.

Example 4 (Nonparametric mean function with GP prior) *We illustrate the use of Gaussian process models in nonparametric regression using a widely studied dataset originally analyzed by Silverman (1985). The data are the measurements of head acceleration in a simulated motorcycle accident used to test crash helmets. The regression function is clearly non-linear, and even a piecewise linear function would have a difficulty fitting this data. The left panel of Figure 2.2 presents the regression function obtained from an interpolation model, corresponding to* $\tau^2 = 0$ *in* (1.9)*. The right panel shows the fit obtained from a smoothing mode* $\tau^2 > 0$*. In both cases, the parameters* τ^2*,* σ^2 *and* λ *where learned from the data using a Markov chain Monte Carlo algorithm.*

Example 5 (Nonparametric regression using wavelets) *Barnes* et al. *(2003) consider data from cepheid stars, i.e., pulsating stars. Figure 2.3 plots observed radial velocities* y_i *against phase* x_i*, together with a non-linear regression estimate* $f(\cdot)$ *based on an BNP model. The model used a basis expansion for the unknown phase-velocity curve. The basis is a wavelet basis. We discussed the prior for this example before, in Example 3. Figure 1.5 shows draws from the prior* $f \sim p(f)$*. We now add a prior to select wavelet coefficients, with* $p(\beta_{j\ell} = 0) = 1 - \alpha^{j+1}$*. Smaller* α *imposes more prior shrinkage and reduces the prior probability for high frequency features. Conditional on the selected wavelet coefficients we continue to use the dependent prior introduced before. Figure 2.3 shows inference under* $\alpha = 0.5$ *and* $\alpha = 0.7$*.*



(a) $\alpha = 0.5$ (b) $\alpha = 0.7$

FIG 2.3. *Phase-velocity curve $f(x)$ for cepheid stars. The figures show the posterior estimated phase-velocity curve $E(f \mid data)$ (thick central line), and pointwise central HPD 50% (light grey) and 95% (dark grey) intervals for $f(x)$. The circles shows the data points. Inference is under an BNP model $p(f)$ using a basis expansion of f with wavelets and $p(\beta_{j\ell} = 0) = 1 - \alpha^{j=1}$. Recall from Example 3 that ρ defines the level of prior dependence.*

2.2.3. Fully Non-Parametric Regression

Finally, one could go non-parametric on both assumptions. We refer to this as a fully nonparametric regression. The sampling model becomes $p(y_i \mid x_i) = G_x$, with a prior on the family of conditional RPMs, $p(G_x,\ x \in X)$. Many commonly used BNP priors for $\mathcal{G} = \{G_x\}$ are variations of dependent DP priors.

Example 6 (Fully nonparametric regression) *Klein and Moeschberger (1997, chapter 1.11) show data from a clinical trial. The data are survival times for patients with tongue cancers. The study investigated the effect of aneuploidy (abnormal number of chromosomes) of the tumor cells. Let $G_x(\cdot)$ denote the distribution of survival times for patients with aneuploid ($x = 1$) and (normal) diploid ($x = 0$) tumor cells. Figure 2.4 shows the Kaplan-Meier estimator of G_x, $x \in \{0, 1\}$, together with an BNP estimate. The BNP estimate is under a DDP prior on $\{G_x,\ x = 0, 1\}$.*

2.3. Mixed Effects Models

BNP priors are often used for model features that are important for appropriate modeling of the observed data, but that are not of interest in themselves. A typical example are random effects distributions in mixed effects models. Random effects are a convenient and common approach to represent the dependence structure in the observed data. Sometimes random effects also have a meaningful interpretation as a property specific to sampling units. For example, when the experimental units are patients in a clinical study, then patient-specific random effects represent the heterogeneity of the patient population, which needs to be accounted for.

In many analyses the distributional assumptions for such random effects distributions are driven entirely by technical convenience and simplicity, using for example a multivariate normal distribution. However, there is often no good scientific reason

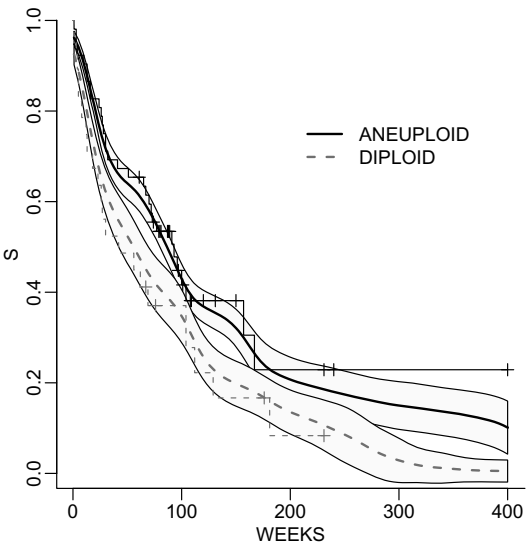


FIG 2.4. *Survival times for tongue cancer patients. The figure shows a Kaplan-Meier estimate (step function) and an NP Bayes estimate (smooth curves) for the survival functions G_x for patients with anaploid ($x = 1$) and diploid ($x = 0$) tumors. The bands around the BNP estimates show pointwise $\pm$ 1.0 posterior standard deviation bounds. The BNP estimate is based on a DDP prior for $\{G_x,\ x = 0, 1\}$.*

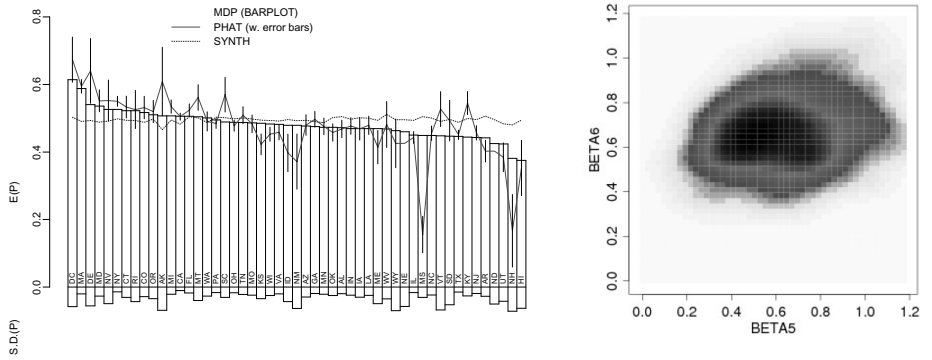
to assume a particular parametric form. Quite to the contrary, patient populations are known to be highly heterogeneous, including outliers, subpopulations and other features that are inconsistent with a multivariate normal model.

This is where BNP priors come in. Let z_i denote a generic random effect specific to the i-th experimental unit. When an investigator wants to avoid a strict parametric assumptions, he or she could instead use $z_i \sim G$ with a BNP prior $G \sim p(G)$. The types of priors used for $p(G)$ are again similar to the density estimation problem, with the difference that in a mixed effects model the random effects z_i are only latent.

Example 7 (Semiparametric mixed effects model) *Malec and Müller (2008) consider a mixed effects model for mammography utilization in the U.S. The data are mammography usage by county and demographic group. The model includes a regression on some county level covariates and county-specific random effects z_i. The random effects are 6-dimensional. Figure 2.5a shows posterior estimated rates of mammography by state. Figure 2.5b shows the estimated random effects distribution.*

2.4. Clustering and Classification

Some statistical inference problems involve a partition of a population of experimental units into clusters. For example, hospitals might be clustered into more homogeneous subgroups, disease subtypes might be grouped by comparable prognosis, states could be grouped by comparable patterns of use of preventive care, etc. Probability models for random partitions can be used to define appropriate



(a) Estimated mammography utilization by state

(b) Estimated r.e. distribution $E(G \mid data)$

FIG 2.5. *Mammography utilization. Estimated rate of mammography utilization by state (a). The histogram shows the estimated use under the semi-parametric Bayesian model. The solid line (with many spikes) shows for comparison the observed sample averages. The dotted line shows an estimate known as synthetic estimate. States are ordered by estimated mammography utilization under the BNP model. Panel (b) shows a bivariate marginal of the estimated random effects distribution $E(G \mid data)$ for county-specific random effects z_i (BETA in the plot).*

inference models for these applications (recall our discussion of product partition models from §1.2.3).

Example 8 (Clustering of morphological data.) *We consider data from Lubischew (1962) who reports measurements on five external characteristics (lengths etc.) of male insects of three species of leaf beetles. We use the 5-dimensional data set, ignoring the species labels. Figure 2.6 shows model-based clustering and classification for this 5-dimensional data set. The plotting symbols show the measurements (showing two of the five dimensions).*

We fit model (2.2) with $\theta_i = (\mu_i, \Sigma_i)$ and $f(x;\, \mu_i, \Sigma_i) = \mathsf{N}(x;\, \mu_i, \Sigma_i)$. The discrete nature of $G \sim \mathsf{DP}$ implies a positive probability of ties among the θ_i. Let $\theta_j^\star$ denote the unique values. The model implies a prior probability model on a partition of the beetles into clusters $S_j = \{i :\ \theta_i = \theta_j^\star\}$ defined by these ties. We will come back to this model several times in the upcoming discussion. The implied prior on the partition of the experimental units, beetles in this case, is known as the Pólya urn. Figure 2.6 shows clusters S_j by different plotting symbols, together with the posterior predictive distribution for a future beetle (contours). In these contours we can recognize the cluster specific $(\mu_j^\star, \Sigma_j^\star)$ as the location $\mu_j^\star$ and orientation $\Sigma_j^\star$ of three ellipses.

2.5. Computation

The flexible nature of BNP inference comes at a price. Implementation of posterior inference for some models can be a bit more involved than similar parametric models. However, actual use of BNP models for data analysis is usually less complicated than what it might seem at first glance. One reason is that inference usually proceeds in a reduced model, after marginalizing with respect to the infinite dimensional quantitity. For example, in a density estimation problem (2.2)

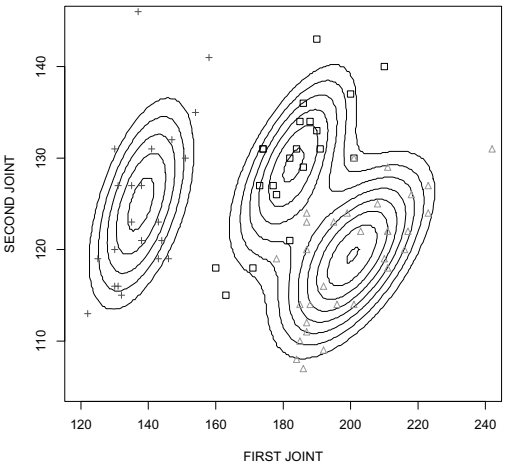


FIG 2.6. *Clustering of beetles by 5 morphological measurements (only two are shown). The plotting symbols show a partition of the beetles into three clusters. The contours show the posterior predictive distribution for a future beetle. We can recognize cluster-specific locations* $\mu_j^\star$ *and covariance matrices* $\Sigma_j^\star$, $j = 1, \ldots, 3$.

with a DP prior, $G \sim \mathsf{DP}(\cdot)$, the marginal model $p(\boldsymbol{\theta}, \boldsymbol{y})$ of all latent variables $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and all data $\boldsymbol{y} = (y_1, \ldots, y_n)$ is available in closed form. This allows for relatively straightforward computation for posterior predictive inference and many other relevant inference summaries.

In Appendix A we show actual implementations in R for inference under DP mixture and PT priors.

Another important feature that makes the use of BNP models practically feasible is the availability of public domain programs. One popular program is the R package `DPpackage` (Jara *et al.*, 2011) that implements inference for PT priors, DP models, Bernstein polynomials, dependent DP models and many variations of these models. The program can be downloaded from `http://www.mat.puc.cl/~ajara/Softwares.html`. The *BNPDensity* package (Barrios *et al.*, 2011) implements density estimation using semi-parametric mixtures with a non-parametric normalized generalized gamma (NGG) prior on the mixing measure (James *et al.*, 2009). Another R package that implements inference for some BNP models is `BayesM` (Rossi *et al.*, 2005).