Nonparametric Bayesian Data Analysis

Peter Müller and Fernando A. Quintana

Abstract. We review the current state of nonparametric Bayesian inference. The discussion follows a list of important statistical inference problems, including density estimation, regression, survival analysis, hierarchical models and model validation. For each inference problem we review relevant nonparametric Bayesian models and approaches including Dirichlet process (DP) models and variations, Pólya trees, wavelet based models, neural network models, spline regression, CART, dependent DP models and model validation with DP and Pólya tree extensions of parametric models.

Key words and phrases: Dirichlet process, regression, density estimation, survival analysis, Pólya tree, random probability model (RPM).

1. INTRODUCTION

Nonparametric Bayesian inference is an oxymoron and a misnomer. Bayesian inference by definition always requires a well-defined probability model for observable data $y$ and any other unknown quantities $\theta$, that is, parameters. Nonparametric Bayesian inference traditionally refers to Bayesian methods that result in inference comparable to classical nonparametric inference, such as kernel density estimation, scatterplot smoothers, etc. Such flexible inference is typically achieved by models with massively many parameters. In fact, a commonly used technical definition of nonparametric Bayesian models is probability models with infinitely many parameters (Bernardo and Smith, 1994). Equivalently, nonparametric Bayesian models are probability models on function spaces. Nonparametric Bayesian models are used to avoid critical dependence on parametric assumptions, to robustify parametric models and to define model diagnostics and sensitivity analysis for parametric models by embedding them in a larger encompassing nonparametric model. The latter two applications are technically simplified by the fact that many nonparametric models allow one to center the probability distribution at a given parametric model.

In this article we review the current state of Bayesian nonparametric inference. The discussion follows a list of important statistical inference problems, including density estimation, regression, survival analysis, hierarchical models and model validation. The list is not exhaustive. In particular, we will not discuss nonparametric Bayesian approaches in time series analysis and in spatial and spatiotemporal inference.

Other recent surveys of nonparametric Bayesian models appear in Walker, Damien, Laud and Smith (1999) and Dey, Müller and Sinha (1998). Nonparametric models based on Dirichlet process mixtures are reviewed in MacEachern and Müller (2000). A recent review of nonparametric Bayesian inference in survival analysis can be found in Sinha and Dey (1997).

2. DENSITY ESTIMATION

The density estimation problem starts with a random sample $x_i \sim F(x_i), i = 1, \ldots, n$, generated from some unknown distribution $F$. A Bayesian approach to this problem requires a probability model for the unknown $F$. Traditional parametric inference considers models that can be indexed by a finite-dimensional parameter, for example, the mean and covariance matrix of a multivariate normal distribution of the appropriate dimension. In many cases, however, constraining inference to a specific parametric form may limit the scope and type of inferences that can be drawn from such models. In contrast, under a nonparametric perspective...
we consider a prior probability model \( p(F) \) for the unknown density \( F \), for \( F \) in some infinite-dimensional function space. This requires the definition of probability measures on a collection of distribution functions. Such probability measures are generically referred to as random probability measures (RPM). Ferguson (1973) states two important desirable properties for this class of measures (see also Antoniak, 1974): (i) their support should be large and (ii) posterior inference should be “analytically manageable.” In the nonparametric context, however, computational aspects are still the subject of much research.

By (ii), in the nonparametric context, computational purposes is the Pólya urn representation by Blackwell and MacQueen (1973). Another very useful result is the construction by Sethuraman (1994): any \( F \sim \mathcal{D}(M, F_0) \) can be represented as

\[
F(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\mu_h}(\cdot), \quad \mu_h \overset{\text{i.i.d.}}{\sim} F_0 \quad \text{and}
\]

\[
w_h = U_h \prod_{j<h} (1 - U_j) \quad \text{with } U_h \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, M).
\]

In words, realizations of the DP can be represented as infinite mixtures of point masses. The locations \( \mu_h \) of the point masses are a sample from \( F_0 \), and the random weights \( w_h \) are generated by a “stick-breaking” procedure. In particular, the DP is an almost surely (a.s.) discrete RPM.

The DP is by far the most popular nonparametric model in the literature (for a recent review, see MacEachern and Müller, 2000). However, the a.s. discreteness is in many applications inappropriate. A simple extension to remove the constraint to discrete measures is to introduce an additional convolution, representing the RPM \( F \) as

\[
F(x) = \int f(x|\theta) dG(\theta) \quad \text{with } G \sim \mathcal{D}(M, G_0).
\]

Such models are known as DP mixtures (MDP) (Escobar, 1988; MacEachern, 1994; Escobar and West, 1995). Using a Gaussian kernel \( f(x|\mu,S) = \phi_{\mu,S}(x) \propto \exp\{- (x - \mu)^T S^{-1} (x - \mu)/2 \} \) and mixing with respect to \( \theta = (\mu,S) \), we obtain density estimates resembling traditional kernel density estimation. Related models have been studied in Lo (1984), Escobar and West (1995) and Gasparini (1996). Posterior consistency is discussed in Ghosal, Ghosh and Ramamoorthi (1999).

Posterior inference in MDP models is based on MCMC posterior simulation. Most approaches proceed by breaking the mixture in (3) with the introduction of latent variables \( \theta_i \) as \( x_i|\theta_i \sim f(x|\theta) \) and \( \theta_i \sim G \). Efficient MCMC simulation for general MDP models is discussed, among others, in Bush and MacEachern (1996), MacEachern and Müller (1998), Neal (2000) and West, Müller and Escobar (1994). For related algorithms in a more general setting, see Ishwaran and James (2001). Alternative to MCMC simulation, sequential importance sampling-based methods have been proposed for MDP models. Examples can be found in Liu (1996), Quintana (1998), MacEachern, Clyde and Liu (1999), Ishwaran and Takahara (2002) and references therein. A third class of methods for MDP models, called the predictive recursion, was proposed by Newton and Zhang (1999). Consider
the posterior predictive distribution in model (3). Let $F_n(B) \equiv E(F(B)|x_1, \ldots, x_n)$ denote the posterior distribution, for any Borel set $B$ in the appropriate space. The Pólya urn representation implies

$$F_1(B) = \frac{M}{M + 1} F_0(B) + \frac{1}{M + 1} P(\theta_1 \in B|x_1).$$

Newton and Zhang (1999) extrapolate this representation to a recursion in the general case:

$$F_i(B) = (1 - w_i) F_{i-1}(B) + w_i P_{i-1}(\theta_i \in B|x_i),$$

where the probability in the second term in the right-hand side of (4) is computed under the current approximation $F_{i-1}$, and the nominal values for the weights are $w_i = 1/(M + i)$, $i \geq 1$. The approximation is exact for $i = 1$. In general, $F_n(B)$ depends on the order in which $x_1, \ldots, x_n$ are processed, but this dependence is rather weak, and in practice it is recommended to average over a number of permutations of the data. The method is very fast to execute and produces very good approximations, although it tends to oversmooth the results. For a comparison of the computational strategies mentioned here, see Quintana and Newton (2000).

Model (1) has the advantage of the conjugate form. However, getting exact draws from a DP is impossible because this requires the generation of an infinite mixture of point masses. Typical MCMC schemes are based on integrating out the DP via Blackwell and MacQueen’s (1973) representation. This makes it difficult to produce inference on functionals of the posterior DP. A similar problem is found in the more general MDP models. Some authors propose MCMC strategies where, instead of integrating out the DP, an approximation to the DP is considered. This is usually done by drawing from $\sum_{h=1}^{N} w_h \delta_{\mu_h}()$ for large enough $N$. Examples of this strategy can be found in Muliere and Tardella (1998), Ishwaran and James (2002), Kottas and Gelfand (2001) and Gelfand and Kottas (2002).

### 2.2 Other Discrete Random Probability Measures

An interesting extension of the DP that has been used in the context of density estimation is the invariant DP introduced by Dalal (1979). The idea is to define a prior process on the space of distribution functions that has a structure that can be characterized via invariance, for example, symmetry or exchangeability.

Dalal’s (1979) construction is based on invariance under a finite group, essentially by restricting Ferguson’s (1973) definition to invariant centering measures and partitions. This guarantees that the posterior process is also invariant. Dalal (1979) uses this setup to estimate distribution functions that are symmetric with respect to a known value $\mu$, using $F_0$ such that $F_0(t) = 1 - F_0(2\mu - t)$ for all $t \leq \mu$ and the group $G = \{g_1, g_2\}$, where $g_1(x) = x$ and $g_2(x) = 2\mu - x$.

An alternative model to (1) or (3) is obtained by replacing the prior DP with a convenient approximation. Natural candidates follow from truncating Sethuraman’s (1994) construction (2). In this setup, the prior $\sum_{h=1}^{N} w_h \delta_{\mu_h}()$ is replaced by $\sum_{h=1}^{N} w_h \delta_{\mu_h}()$ for some appropriately chosen value of $N$. An example of this procedure is the $\epsilon$-DP proposed by Muliere and Tardella (1998), where $N$ is chosen such that the total variation distance between the DP and the truncation is bounded by a given $\epsilon$. Another variation is the Dirichlet-multinomial process introduced by Muliere and Secchi (1995). Here the RPM is, for some finite $N$,

$$F(\cdot) = \sum_{h=1}^{N} w_h \delta_{\mu_h}(),$$

$$(w_1, \ldots, w_N) \sim D(M \cdot N^{-1}, \ldots, M \cdot N^{-1})$$

and $\mu_h \sim F_0$. More generally, Pitman (1996) described a class of models

$$F(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\mu_h}() + \left(1 - \sum_{h=1}^{\infty} w_h\right) F_0(\cdot),$$

where, for a continuous distribution $F_0$, we have $\mu_h \overset{i.i.d.}{\sim} F_0$, assumed independent of the nonnegative random variables $w_h$. The weights $w_h$ are constrained by $\sum_{h=1}^{\infty} w_h \leq 1$. The model is known as a species sampling model (SSM), with $w_h$ interpreted as the relative frequency of the $h$th species in a list of species present in a certain population, and $\mu_h$ as the tag assigned to that species. If $\sum_{h=1}^{\infty} w_h = 1$, the SSM is called proper and the corresponding prior RPM is discrete. The stick-breaking priors studied by Ishwaran and James (2001) are a special case of (5), adopting the form $\sum_{h=1}^{N} w_h \delta_{\mu_h}()$, where $1 \leq N \leq \infty$. The weights are defined as $w_h = \prod_{j=1}^{h-1} (1 - U_j) U_h$ with $U_h \sim Beta(a_h, b_h)$, independently, for given sequences $(a_1, a_2, \ldots)$ and $(b_1, b_2, \ldots)$. Stick-breaking priors are quite general, including not only the Dirichlet-multinomial process and the DP as special cases, but
also a two-parameter DP extension, known as the Pitman–Yor process (Pitman and Yor, 1997), and the Beta two-parameter process (Ishwaran and Zarepour, 2000). Additional examples and MCMC implementation details for stick-breaking RPM’s can be found in Ishwaran and James (2001). Further discussion of SSM’s appears in Pitman (1996) and Ishwaran and James (2003).

An interesting property of MDP models is that any exchangeable sequence of random variables can be well approximated in the sense of the Prokhorov metric by a certain sequence of mixtures of DP’s (Regazzini, 1999). In practice, however, this result has limited use. We review next some methods for defining RPM’s supported on the set of continuous distributions that have been used in density estimation problems.

2.3 Pólya Trees

Pólya trees (PT) are proposed in Lavine (1992, 1994) as a generalization of the DP. Like the DP, the PT model satisfies conditions (i) and (ii). The PT includes DP models as a special case. However, in contrast to the DP, an appropriate choice of the PT model satisfies conditions (i) and (ii). The PT includes DP models as a special case. However, in contrast to the DP, an appropriate choice of the PT model satisfies conditions (i) and (ii). The PT model satisfies conditions (i) and (ii). The PT model satisfies conditions (i) and (ii). The PT model satisfies conditions (i) and (ii).

The type of models used for density estimation now replaces the DP in (1) and (3) by the PT. An interesting property of MDP models is that any exchangeable sequence of random variables can be well approximated in the sense of the Prokhorov metric by a certain sequence of mixtures of DP’s (Regazzini, 1999). In practice, however, this result has limited use. We review next some methods for defining RPM’s supported on the set of continuous distributions that have been used in density estimation problems.

2.4 Bernstein Polynomials

For a distribution function \( F \) on the unit interval, the corresponding Bernstein polynomial is defined as

\[
B(x, k, F) = \sum_{j=0}^{k} \binom{k}{j} x^j (1-x)^{k-j}.
\]

A remarkable property of \( B(x, k, F) \) is that it converges uniformly to \( F \) as \( k \to \infty \). The definition for \( B(x, k, F) \) takes the form of a mixture of Beta densities. Petrone (1999a, b) exploits this property to propose a class of prior distributions on the set of densities defined on \((0, 1)\). Petrone and Wasserman (2002) consider the following model. Assume \( x_1, \ldots, x_n \) are conditionally i.i.d. given \( k \) and \( w_k \) with common density

\[
f(x|k, w_k) = \sum_{j=1}^{k} w_{jk} \frac{k!}{(j-1)!} x^{j-1} (1-x)^{k-j},
\]

where \( k \) is the number of components in the mixture of Beta densities and the weights \( w_k = (w_{1k}, \ldots, w_{kk}) \) satisfy \( w_{jk} \geq 0 \) and \( \sum_{j=1}^{k} w_{jk} = 1 \). We call \( f \) a Bernstein polynomial density (BPD). The model is completed by assuming a prior distribution \( p(k) \) for \( k \) and a distribution \( H_k(\cdot) \) given \( k \) on the \((k-1)\)-dimensional simplex. Petrone (1999a) showed that if \( p(k) > 0 \) for all \( k \geq 1 \), then every distribution on \((0, 1)\) is the (weak) limit of some sequence of BPD, and every continuous density on \((0, 1)\) can be well approximated in the Kolmogorov–Smirnov distance by BPD. Petrone and Wasserman (2002) discuss MCMC strategies for fitting the above model and prove consistency of posterior density estimation under mild conditions. Rates of such convergence are given in Ghosal (2001).
2.5 Other Random Distributions

Lenk (1988) introduces the logistic normal process. The construction of a logistic normal process starts with a Gaussian process $Z(x)$ with mean function $\mu(x)$ and covariance function $\sigma(x, y)$. The transformed process $W = \exp(Z)$ is a lognormal process. Stopping the construction here and defining a random density $f(x) \propto W$ would be impractical. The lognormal process is not closed under prior to posterior updating; that is, the posterior on $f$ results in a lognormal process. Instead Lenk (1988) proceeds by defining the logistic normal process. Instead Lenk (1988) proceeds by defining the generalized lognormal process, essentially by weighting realizations under the lognormal process with the random integral $(\int W \, d\lambda)^{\varepsilon}$. Let $f(x) \propto V(x)$ for $V \sim LN_X(\mu, \sigma, \xi)$. The density $f$ is said to be a logistic normal process $LNS_X(\mu, \sigma, \xi)$. The posterior on $f$, conditional on a random sample $y \sim f$, is again a logistic normal process $LNS_X(\mu^*, \sigma, \xi^*)$.

The updated parameters are $\mu^*(s) = \mu(s) + \sigma(s, y)$ and $\xi^* = \xi - 1$.

3. REGRESSION

The generic regression problem seeks to estimate an unknown mean function $g(x)$ based on data with i.i.d. measurement errors: $y_i = g(x_i) + \varepsilon_i$, $i = 1, \ldots, n$. Bayesian inference on $g$ starts with a prior probability model for the unknown function $g$. If restrictive parametric assumptions for $g$ are inappropriate, we are led to consider nonparametric Bayesian models. Many approaches proceed by considering some basis $B = \{f_1, f_2, f_3, \ldots\}$ for an appropriate function space, like the space of square integrable functions. Typical examples are the Fourier basis, wavelet bases and spline bases. Given a chosen basis $B$, any function $g$ can be represented as $g(\cdot) = \sum b_i f_i(\cdot)$. A random function $g$ is parametrized by the sequence $b = (b_1, b_2, \ldots)$ of basis coefficients. Assuming a prior probability model for $b$ we implicitly put a prior probability model on the random function.

3.1 Spline Models

A commonly used class of basis functions are splines, for example, cubic regression splines $B = \{1, x, x^2, x^3, (x - \xi_1)^3, \ldots, (x - \xi_T)^3\}$, where $(x)_+ = \max(x, 0)$ and $\xi = (\xi_1, \ldots, \xi_T)$ is a set of knots. Together with a normal measurement error $\varepsilon_i \sim N(0, \sigma)$ this defines a nonparametric regression model

\begin{equation}
\hat{y}_i = \sum b_i f_i(x_i) + \varepsilon_i.
\end{equation}

The model is completed with a prior $p(\xi, b, \sigma)$ on the set of knots and corresponding coefficients. Smith and Kohn (1996), Denison, Mallick and Smith (1998b) and DiMatteo, Genovese and Kass (2001) are typical examples of such models. Approaches differ mainly in the choice of priors and the implementation. Typically the prior is assumed to factor, $p(\xi, b, \sigma) = p(\xi)p(\sigma)p(b|\sigma)$. Smith and Kohn (1996) use the Zellner g-prior (Zellner, 1986) for $p(b)$. The prior covariance matrix $\text{Var}(b|\sigma)$ is assumed to be proportional to $(B'B)^{-1}$, where $B$ is the design matrix for the given data set. Assuming a conjugate normal prior $b \sim N(0, c\sigma(B'B)^{-1})$, the conditional posterior mean $E(b|\xi, \sigma)$ is a simple linear shrinkage of the least squares estimate $\hat{b}$. DiMatteo, Genovese and Kass (2001) use a unit-information prior which is defined as a Zellner g-prior with the scalar $c$ chosen such that the prior variance is equivalent to one observation. Denison, Mallick and Smith (1998b) prefer a ridge prior $p(b) = N(0, V)$ with $V = \text{diag}(\alpha, v, \ldots, v)$.

Posterior simulation in (6) is straightforward except for the computational challenge of updating $\xi$, the number and location of knots. This typically involves reversible jump MCMC (Green, 1995). Denison, Mallick and Smith (1998a) propose “birth,” “death” and “move” proposals to add, delete and change knots from the currently imputed set $\xi$ of knots. In the implementation of these moves it is important to marginalize with respect to the coefficients $b_i$.

In the conditionally conjugate setup with a normal prior $p(b|\sigma)$ the marginal posterior $p(\xi|\sigma, y)$ can be evaluated analytically. DiMatteo, Genovese and Kass (2001) propose an approximate evaluation of the relevant Bayes factors based on the Bayesian information criterion (BIC). An interesting alternative, called focused sampling, is discussed in Smith and Kohn (1998).

3.2 Multivariate Regression

Extensions of spline regression to multiple covariates are complicated by the curse of dimensionality. Smith and Kohn (1997) define a spline based bivariate regression model. General, higher-dimensional regression models require some simplifying assumptions about the nature of interactions to allow a practical implementation. One approach is to assume additive effects

\begin{equation}
\hat{y}_i = \sum_j g_j(x_{ij}) + \varepsilon_i,
\end{equation}

and proceed with each $g_j$ as before. Shively, Kohn and Wood (1999) and Denison, Mallick and Smith (1998b)
propose such implementations. Denison, Mallick and Smith (1998c) explore an alternative extension of univariate splines, following the idea of multivariate adaptive regression splines (MARS; Friedman, 1991). MARS uses basis functions that are constructed as products of univariate functions. Let \( x_i = (x_{i1}, \ldots, x_{ip}) \) denote the multivariate covariate vector. MARS assumes
\[
g(x_i) = b_0 + \sum_{h=1}^{k} b_h f_h(x_i)
\]
with
\[
f_h(x) = \prod_{j=1}^{J_h} s_h j(x w_{hj} - t_{hj})_+.
\]
Here we used linear spline terms \((x - t_{hj})_+\) to construct the basis functions \( f_h \). Each basis function defines an interaction of \( J_h \) covariates. The indices \( w_{hj} \) specify the covariates and the \( t_{hj} \) give the corresponding knots.

Another intuitively appealing multivariate extension is classification and regression tree (CART) models. Chipman, George and McCulloch (1998) and Denison, Mallick and Smith (1998a) discuss Bayesian inference in CART models. A regression tree is parametrized by Mallick and Smith (1998a) discuss Bayesian inference

Chipman, George and McCulloch (1998) and Denison, Mallick and Smith (1998c) explore an alternative extension of classification and regression tree (CART) models. At the top level of a hierarchical model they assume a Poisson process as a sampling model for the observed data. Let \( x_i \) denote the coordinates of an observed event. For example, \( x_i \) could be the recorded occurrence of a species in a species sampling problem. The model assumes a Poisson process \( x_i \sim \text{Po}(\Lambda(x)) \) with intensity function \( \Lambda(x) \). The intensity function in turn is modeled as a convolution of a normal kernel \( k(x, s) \) and a Gamma process, \( \Lambda(x) = \int k(x, s) \Gamma(ds) \) and \( \Gamma(ds) \sim \text{Gamma}(\alpha(ds), \beta(ds)) \). With constant \( \beta(s) = \beta \) and rescaling the Gamma process to total mass 1, the model for \( \Lambda(x) \) reduces to a Dirichlet process mixture of normals.

Arjas and Heikkinen (1997) propose an alternative approach to inference for a spatial Poisson process. The prior probability model is based on Voronoi tessellations with a random number and location of knots.

### 3.3 Wavelet Based Modeling

Wavelets provide an orthonormal basis in \( L^2 \) representing \( g \in L^2 \) as \( g(x) = \sum_{j} \sum_{k} d_{jk} \psi_{jk}(x) \), with basis functions \( \psi_{jk}(x) = 2^{j/2} \psi(2^j x - k) \) that can be expressed as shifted and scaled versions of one underlying function \( \psi \). The practical attraction of wavelet bases is the availability of superfast algorithms to compute the coefficients \( d_{jk} \) given a function, and vice versa. Assuming a prior probability model for the coefficients \( d_{jk} \) implicitly puts a prior probability model on the random function \( g \). Typical prior probability models for wavelet coefficients include positive probability mass at zero. Usually this prior probability mass depends on the “level of detail” \( j \), \( \Pr(d_{jk} = 0) = \pi_j \).
Given a nonzero coefficient, an independent prior with level dependent variances is assumed, for example, \( p(d_{jk}|d_{jk} \neq 0) = N(0, \tau_j^2) \). Appropriate choice of \( \tau_j \) and \( \tau_j \) achieves posterior rules for the wavelet coefficients \( d_{jk} \), which closely mimic the usual wavelet thresholding and shrinkage rules (Chipman, Kolaczyk and McCulloch, 1997; Vidakovic, 1998). Clyde and George (2000) discuss the use of empirical Bayes estimates for the hyperparameters in such models.

Posterior inference is greatly simplified by the orthonormality of the wavelet basis. Consider a regression model \( y_i = g(x_i) + \varepsilon_i, i = 1, \ldots, n \), with equally spaced data \( x_i \), for example, \( x_i = i/n \). Substitute a wavelet basis representation \( g(\cdot) = \sum_j \sum_k d_{jk} \psi_{jk}(x) \), and let \( y, d \) and \( \varepsilon \) denote the data vector, the vector of all wavelet coefficients and the residual vector, respectively. Also, let \( B = [\psi_{jk}(x_i)] \) denote the design matrix of the wavelet basis functions evaluated at the \( x_i \). Then we can write the regression in matrix notation as \( y = Bd + \varepsilon \). The discrete wavelet transform of the data finds, in a computationally highly efficient algorithm, \( \hat{d} = B^{-1}y \). Assuming independent normal errors, \( \varepsilon_i \sim N(0, \sigma^2) \), orthogonality of the design matrix \( B \) implies \( \hat{d}_{jk} \sim N(d_{jk}, \sigma^2) \), independently across \( (j, k) \). Assuming a priori independent \( d_{jk} \) leads to a posteriori independence of the wavelet coefficients \( d_{jk} \).

In other words, we can consider one univariate inference problem \( p(d_{jk}|y) \) at a time. Even if the prior probability model \( p(d) \) is not marginally independent across \( d_{jk} \), it typically assumes independence conditional on hyperparameters, still leaving a considerable simplification of posterior simulation.

The above detailed explanation serves to highlight two critical assumptions. Posterior independence, conditional on hyperparameters or marginally, only holds for equally spaced data and under a priori independence over \( d_{jk} \). In most applications prior independence is a technically convenient assumption, but does not reflect genuine prior knowledge. However, incorporating assumptions about prior dependence is not excessively difficult either. Starting with an assumption about dependence of \( g(x_i) \), \( i = 1, \ldots, n \), Vannucci and Corradi (1999) show that a straightforward two-dimensional wavelet transform can be used to derive the corresponding covariance matrix for the wavelet coefficients \( d_{jk} \).

In the absence of equally spaced data the convenient mapping of the raw data \( y_i \) to the empirical wavelet coefficients \( \hat{d}_{jk} \) is lost. The same is true for inference problems other than regression where wavelet decomposition is used to model random functions. Typical examples are the unknown density in a density estimation (Müller and Vidakovic, 1998) and the spectrum in a spectral density estimation (Müller and Vidakovic, 1999). In either case evaluation of the likelihood \( p(y|d) \) requires reconstruction of the random function \( g(\cdot) \). Although a technical inconvenience, this does not hinder the practical use of a wavelet basis. The superfast wavelet decomposition and reconstruction algorithms still allow computationally efficient likelihood evaluation even with the original raw data.

### 3.4 Neural Networks

Neural networks are another popular approach following the general theme of defining random functions by probability models for coefficients with respect to an appropriate basis. Now the bases are rescaled versions of logistic functions. Let \( \Psi(\eta) = \exp(\eta)/(1 + \exp(\eta)) \); then \( g(x) = \sum_{j=1}^{M} \beta_j \Psi(x'\gamma_j) \) can be used to represent a random function \( g \). The random function is parameterized by \( \theta = (\beta_1, \gamma_1, \ldots, \beta_M, \gamma_M) \). Bayesian inference proceeds by assuming an appropriate prior probability model and considering posterior updating conditional on the observed data. Recent reviews of statistical inference for neural networks in regression models appear in Cheng and Titterington (1994) and Stern (1996). Neal (1996) and Müller and Ríos-Insua (1998) discuss specifically Bayesian inference in such models. Ríos-Insua and Müller (1998) argue to include the number of components \( M \) in the parameter vector and consider inference over “variable architecture” neural network models. Lee (2001) compares alternative Bayesian model selection criteria for neural networks.

### 3.5 Other Nonparametric Regression Methods

Alternatively to modeling the random function \( g \), the nonparametric regression problem can be reduced to a density estimation problem by proceeding as if the pairs \((x_i, y_i)\) were an i.i.d. sample, \((x_i, y_i) \sim F(x, y)\), from some unknown distribution \( F \). Inference on \( F \) implies inference on the conditional means process \( g_F(x) \equiv E_F(y|x) \). Müller, Erkanli and West (1996) propose this approach using a DP mixture model for inference on the unknown joint distribution \( F \). Regression curves \( g \) estimated under this approach take the form of locally weighted linear regression lines, similar to traditional kernel regression in classical nonparametric inference. Considering \((x_i, y_i)\) as an i.i.d. sample—wrongly—introduces an addi-
tional factor $\prod F(x_i)$ in the likelihood $\prod F(x_i, y_i) = \prod F(x_i) F(y_i | x_i)$ and thus provides only approximate inference.

An interesting approach to isotonic regression is pursued in Lavine and Mockus (1995), who use a rescaled cumulative density function $F$ to model a regression mean curve $g(x) = a + bF(x)$. Assuming a DP prior for $F$ they implement nonparametric Bayesian inference.

Newton, Czado and Chappell (1996) propose a modified DP, constraining the random probability measure to median 0 and fixed length central interval (such as, e.g., the interquartile range). The modified DP is used to define a link $F$ in a nonparametric binary regression model with $P(y_i = 1) = F(x_i^\beta)$.

4. SURVIVAL ANALYSIS

Survival analysis involves modeling the time until a certain event occurs (survival times), often including a regression on covariates. In most applications, the data is subject to right-censoring. Let $x_1, \ldots, x_n$ denote the survival times, $x_i \sim F(\cdot)$. Let $C_1, \ldots, C_n$ denote the (possibly random) censoring times. The actually observed data is a collection of pairs $(T_1, I_1), \ldots, (T_n, I_n)$ with censored observations $T_i = \min\{x_i, C_i\}$ and censoring indicators $I_i = I\{x_i \leq C_i\}$. Interval and other types of censoring could be also considered in a similar fashion. Two quantities are of primary interest in survival analysis: the survival function $S(t) = 1 - F(t)$ and the hazard rate function $\lambda(t) = F'(t)/S(t)$.

It turns out that the integrated or cumulative hazard function $\Lambda(t) = \int_{0}^{t} \lambda(s) ds$ is simpler to estimate, and there is a one-to-one correspondence between $S(t)$ and $\Lambda(t)$, given by $S(t) = \exp(-\Lambda(t))$.

Assuming $C_1, \ldots, C_n$ to be constant, Susarla and Van Ryzin (1976) discuss inference with a DP prior on $F$. The posterior mean converges to Kaplan and Meier’s (1958) product limit estimate as the total mass parameter $M \rightarrow 0^+$. More recently, Florens and Rolin (2001) provided a closed form description of the posterior process under a DP prior and random censoring times. The characterization is quite useful for posterior simulation of functionals of the posterior distribution of $F$. For a review of related approaches applying the DP to similar problems see Ferguson, Phadia and Tiwari (1992). Doss (1994) studied an MDP model for survival data subject to more general censoring schemes. Evaluation of the posterior mean of $F$ is done through an interesting MCMC scheme that involves DP draws using a composition method. Convergence of the algorithm is also discussed.

4.1 Neutral to the Right Processes

Many stochastic process priors that have been proposed as nonparametric prior distributions for survival data analysis belong to the class of neutral to the right (NTTR) processes. An RPM $F(t)$ is an NTTR process on the real line if it can be expressed as $F(t) = 1 - \exp(-Y(t))$, where $Y(t)$ is a stochastic process with independent increments, almost surely right-continuous and nondecreasing with $P\{Y(0) = 0\} = 1$ and $P\{\lim_{t \rightarrow \infty} Y(t) = \infty\} = 1$ (Doksum, 1974). Walker et al. (1999) call $Y(t)$ an NTTR Lévy process. Doksum (1974) showed that the posterior for an NTTR prior and i.i.d. sampling is again an NTTR process. Ferguson and Phadia (1979) showed that for right-censored data the class of NTTR process priors remains closed; that is, the posterior is still an NTTR process.

NTTR processes are used in many approaches that construct probability models for $\lambda(t)$ or $\Lambda(t)$, rather than directly for $F$. Dykstra and Laud (1981) define the extended Gamma process, generalizing the Gamma process studied in Ferguson (1973). The idea is to consider first an NTTR process $\{Y(t)\}$ such that $Y(t) - Y(s) \sim \Gamma(\alpha(t) - \alpha(s), 1)$ for all $t > s \geq 0$, where $\alpha(t)$ is a nondecreasing left-continuous function on $[0, \infty)$. The new process is defined as $f_{\alpha} \beta(s) dY(s)$ for a positive right-continuous function $\beta(t)$. Dykstra and Laud (1981) consider such processes on the hazard function $\lambda(t)$, studying their properties and obtaining estimates of the posterior hazard function without censoring and with right-censoring. In particular, the resulting function $\lambda(t)$ is monotone.

An alternative model was proposed by Hjort (1990), by placing a Beta process prior on $\Lambda(t)$. To understand this construction, let us look at a discrete version of the process first. Following Nieto-Barajas and Walker (2002b), consider a partition of the time axis $0 = t_0 < t_1 < t_2 < \cdots$, and failures occurring at times chosen from the set $\{t_1, t_2, \ldots\}$. Let $\lambda_j$ denote the hazard at time $t_j$, $\lambda_j = P(x = t_j | x \geq t_j)$. Hjort (1990) assumes independent, Beta-distributed priors for $\{\lambda_j\}$. This generates a discrete process with independent increments for the cumulative hazard function $\Lambda(t_j) = \sum_{i=0}^{j} \lambda_i$. The class is closed under prior to posterior updating as the posterior process is again of the same type. The continuous version of this discrete Beta process is derived by a limit argument as the interval lengths $t_j - t_{j-1}$ approach zero (for details, see Hjort, 1990). Full Bayesian inference for a model with a Beta process prior for the cumulative hazard function using Gibbs sampling can be found in Damien,
Thus, in general the method leads to an approximate simulation. The name ILM comes from the fact that $v_m = L^{-1}(\tau_m / T)$, where $L(u) = v([u, \infty))$. See additional details in Wolpert and Ickstadt (1998b).

4.2 Dependent Increments Models

We have already discussed independent increments models for the cumulative hazard function $\Lambda(t)$. In the discrete version this implies independence for the hazards $\Lambda_j$. A different modeling perspective is obtained by assuming dependence. A convenient way to introduce dependence is a Markovian process prior on $\lambda_k$. Gamerman (1991) proposes the following model: $\log \lambda_j = \log \lambda_{j-1} + \epsilon_j$ for $j \geq 2$, where $\epsilon_j$ are independent with $E(\epsilon_j) = 0$ and $\text{Var}(\epsilon_j) = \sigma^2 < \infty$. In the linear Bayesian method of Gamerman (1991) only a partial specification of the $\epsilon_j$ is required. The resulting model extends Leonard’s (1978) smoothness prior for density estimation, stated also in terms of a discrete survival formulation, but under the assumption that $\epsilon_j \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

Later, Gray (1994) used a similar prior process but directly on the hazards $\lambda_j$, without the log transformation. A further generalization involving a martingale process was proposed in Arjas and Gasbarra (1994). More recently, Nieto-Barajas and Walker (2002b) proposed a model based on a latent process $\{u_k\}$ such that $\lambda_j$ is included as

$$\lambda_1 \rightarrow u_1 \rightarrow \lambda_2 \rightarrow u_2 \rightarrow \cdots$$

and the pairs $(u, \lambda)$ are generated from conditional densities $f(u|\lambda)$ and $f(\lambda|u)$ implied by a specified joint density $f(u, \lambda)$. The main idea is to ensure linearity in the conditional expectation: $E(\lambda_{k+1}|\lambda_k = a_k + b_k \lambda_k$. Nieto-Barajas and Walker (2002b) show that both the Gamma process of Walker and Mallick (1997) and the discrete Beta process of Hjort (1990) are obtained as special cases of their construction, under appropriate choices of $f(u, \lambda)$.

In the continuous case, Nieto-Barajas and Walker (2002b) proposed a Markovian model where the hazard rate function is modelled as

$$\lambda(t) = \int_0^t \exp(-a(t - u)) \, dL(u),$$

for $a > 0$, and where $L(t)$ is a pure jump process, that is, an independent increments process on $[0, \infty)$ without Gaussian components (Ferguson and Klass, 1972; Walker and Damien, 2000). This model, called a Lévy driven Markov process, extends Dykstra and Laud’s (1981) proposal by allowing nonmonotone sample
paths for $\lambda(t)$. In addition, the sample paths are piecewise continuous functions. Nieto-Barajas and Walker (2002b) obtain posterior distributions under (7) for different types of censoring and discuss applications in several special cases, including the Markov–Gamma process.

### 4.3 Competing Risks Model

An interesting extension of survival models considers a system with $r$ components arranged in series. Here $x_1, \ldots, x_r$ are the failure times of the components and we observe $(T, I)$, where $T = \min\{x_1, \ldots, x_r\}$ and $I = j$ if $T = x_j$. This setup is known as the competing risks model with $r$ sources of failure. The survival function for the $j$th component is

$$S_j(t) = P(x_j > t)$$

and the subsurvival function is

$$S^*_j(t) = P(T > t, I = j).$$

The system survival function is

$$S(t) = P(T > t) = \sum_{j=1}^{r} S^*_j(t).$$

Let $x_i = (x_{i1}, \ldots, x_{ir})$, $i = 1, \ldots, n$, be a sample from the latent $x_1, \ldots, x_r$ failure times. The actual observed data are $(T_1, I_1), \ldots, (T_n, I_n)$. Salinas-Torres, de Bragança Pereira and Tiwari 1997 introduced the multivariate DP as a nonparametric model for the joint distribution of the failure times $x_1, \ldots, x_n$. Let $F_0, \ldots, F_r$ be distribution functions on the appropriate space and $M_1, \ldots, M_r$ be positive mass parameters, and let $v = (v_1, \ldots, v_r) \sim D(M_1, \ldots, M_r)$. Then $P = (v_1 P_1, \ldots, v_r P_r)$ is called a multivariate DP of dimension $r$ if $P_j \sim D(M_j, F_0)$.

Consider now a given risk subset $\Delta \subset \{1, \ldots, r\}$ and let $\Delta^c$ be its complement. The corresponding sub-subsurvival and survival functions are given by $S^*_\Delta(t) = P(T > t, I \in \Delta)$ and $S_\Delta(t) = P(\min_{j \in \Delta} x_j > t)$. The data structure obtained for the case $r = 2$, $\Delta = \{1\}$ and $\Delta^c = \{2\}$ reduces to the usual right-censored problem with random censoring times. Peterson (1977) gives an expression for the survival function $S_\Delta(t)$ in terms of the subsurvival functions $S^*_\Delta(t)$ and $S^*_{\Delta^c}$:

$$S_\Delta(t) = \varphi(S^*_\Delta(t), S^*_{\Delta^c}; t)$$

where

$$\varphi(H, G; t) = \exp\left(\phi^{t_0}_t \frac{dH(s)}{H(s) + G(s)} \prod_{i} \frac{H(s_+)}{H(s_-)} + G(s_+)\right) \prod_{i} \frac{H(s_-)}{H(s_+) + G(s_-)},$$

and $t_{S_\Delta} = \sup\{t : S_\Delta(t) > 0\}$. Here, $\phi^{t_0}_t$ represents integration over the union of intervals of continuity points of $H$ that are less than or equal to $t$, and $\prod_{i}$ represents a product over the discontinuity points of $H$ that are less than or equal to $t$ [we assume that $S^*_\Delta(t)$ and $S^*_{\Delta^c}(t)$ have no common discontinuities]. In this setting, Salinas-Torres, de Bragança Pereira and Tiwari (2002) derived Bayes estimates of $S_\Delta(t)$ under quadratic loss function. The estimate has the property that it can be obtained by substituting the Bayes estimate for $S^*_\Delta$ and $S^*_{\Delta^c}$ into (8).

### 4.4 Models Based on Proportional Hazards

So far we have discussed survival analysis models without covariates. To incorporate covariates, the most popular choice is the proportional hazards model, introduced in Cox (1972). Assuming $T_1, \ldots, T_n$ are the failure times of $n$ individuals, the hazard rate functions are modeled as

$$\lambda_i(t) = \lambda_0(t) \exp[Z_i(t)^T \beta], \quad i = 1, \ldots, n,$$

where $Z_i(t)$ is the $p$-dimensional vector of covariates for the $i$th individual at time $t > 0$, $\beta$ is the vector of regression coefficients and $\lambda_0(t)$ is the baseline hazard rate function.

Semiparametric approaches to inference in (9) consider a nonparametric specification of $\lambda_0(t)$. A model based on independent increments Gamma process was proposed by Kalbfleisch (1978), who studied its properties and estimation. Extensions of this model to neutral-to-the-right processes were discussed in Wild and Kalbfleisch (1981). In the context of multiple event time data, Sinha (1993) considered an extension of Kalbfleisch’s (1978) model for $\lambda_0(t)$. The proposal assumes the events are generated by a counting process with intensity given by a multiplicative expression similar to (9), but including an indicator of the censoring process, and individual frailties to accommodate the multiple events occurring per subject. Sinha (1993) discusses posterior inference for this model using Gibbs sampling, under the assumption of Gamma-distributed frailties. Extensions of this model to the case of positive stable frailty distributions and a correlated prior process with piecewise exponential hazards can be found in Qiou, Ravishanker and Dey (1999). See additional comments, details on computational strategies and extensions to multivariate survival data in Sinha and Dey (1998).

Other modeling approaches based on (9) have been studied in the literature. Laud, Damien and Smith (1998) consider (9) using a Beta process prior for $\Lambda(t)$,
and proposing an MCMC implementation for full posterior inference. Nieto-Barajas and Walker (2001) propose their flexible Lévy driven Markov process (Nieto-Barajas and Walker, 2002a) to model $\lambda_i(t)$, and allow for time dependent covariates. Full posterior inference is achieved via substitution sampling.

Accelerated failure time models are an alternative framework to introduce regression in survival analysis. Instead of introducing the regression in the log hazard, as in (9), the generic accelerated failure time model assumes that failure times $T_i$ arise as $\log T_i = -z_i^T \beta + \log(x_i)$. Nonparametric approaches assume a probability model for the unknown distribution of $\log(x_i)$. Models based on DP priors appear in Johnson (1985), and MacEachern (1996) propose a DP prior for the distribution of $\log(x_i)$. Models based on DP priors appear in Johnson and Christensen (1989) and Kuo and Mallick (1997). Walker and Mallick (1999) propose an alternative PT prior model.

### 5. HIERARCHICAL MODELS

An important application of nonparametric approaches arises in modeling random effects distributions in hierarchical models. Often little is known about the specific form of the random effects distribution. Assuming a specific parametric form is typically motivated by technical convenience rather than by genuine prior beliefs. Although inference about the random effects distribution itself is rarely of interest, it can have implications for the inference of interest. Thus it is important to allow for population heterogeneity, outliers, skewness, etc.

In the context of a traditional randomized block ANOVA model with subject specific random effects $z_i$, a Bayesian nonparametric model can be used to allow for more general random effects distributions. Bush and MacEachern (1996) propose a DP prior for $z_i \sim G$, $G \sim \mathcal{D}(G_0, M)$. Kleinman and Ibrahim (1998) propose the same approach in a more general framework for a linear model with random effects. They discuss an application to longitudinal random effects models. Müller and Rosner (1997) use DP mixture of normals to avoid the awkward discreteness of the implied random effects distribution. Also, the additional convolution with a normal kernel significantly simplifies posterior simulation for sampling distributions beyond the normal linear model. Mukhopadhyay and Gelfand (1997) implement the same approach in generalized linear models with linear predictor $z_i + x_i^T \beta$ and a DP mixture model for the random effect $z_i$. In Wang and Taylor (2001) random effects $W_i$ are entire longitudinal paths for each subject in the study. They use integrated Ornstein–Uhlenbeck stochastic process priors for $W_i S(t)$.

A further complication arises when the model hierarchy in a hierarchical model continues beyond the nonparametric model, that is, if the nonparametric model appears in a submodel of the larger hierarchical model. For example, in a hierarchical analysis of related clinical studies there might be a different random effects distribution in each of the related clinical trials. Let $G_i$ denote the random distribution or random function in submodel $i$. Assuming a nonparametric model $p(G_i)$ for the $i$th submodel, model completion requires an additional assumption about the joint distribution of $\{G_i, i \in I\}$. Using DP priors, $G_i \sim \mathcal{D}(G_0^{i}, M)$, marginally for each $G_i$, a conceptually straightforward approach is to link the base measures $G_0^{i}$. For example, the base measure $G_0^{i}$ could include a regression on covariates specific to the $i$th submodel. This construction is introduced in Cifarelli and Regazzini (1978) as mixture of products of Dirichlet process. The model is used, for example, in Muliere and Petrone (1993), who define dependent nonparametric models $F_x \sim \mathcal{D}(M, F^0_x)$ by assuming a regression in the base measure $F^0_x \sim N(\beta x, \sigma^2)$. Similar models are discussed in Mira and Petrone (1996) and Giudici, Mezzetti and Muliere (2003). Caroata and Parmigiani (2002) and Dominici and Parmigiani (2001) use the same approach to model random distributions $G_i \sim \mathcal{D}(G_0^{i}, M_i)$ centered around, among other choices, a Binomial base measure $G_0^{i} = \text{Bin}(\theta_0^{i}, N_i)$, including the total mass parameter $M_i$ in the hierarchy. Both the Binomial success probability $\theta_0^{i}$ and the total mass parameter $M_i$ are modeled as a regression on covariates $d_i$, specific to submodel $i$.

Linking the related nonparametric models through a regression on the parameters of the nonparametric models limits the nature of the dependence to the structure of this regression. MacEachern (1999) proposes the dependent DP (DDP) as an alternative approach to define a dependent prior model for a set of random measures $\{G_x\}$, with $G_x \sim \mathcal{D}$ marginally. Recall the stick-breaking representation (2) for the DP random measure, $G_x = \sum_h w_{xh} \delta(\mu_{xh})$. The key idea behind the DDP is to introduce dependence across the measures $G_x$ by assuming the distribution of the point masses $\mu_{xh}$ to be dependent across different levels of $x$, but still independent across $h$. In the basic version of the DDP the weights are assumed to be the same across $x$, that is, $w_{xh} = w_h$. To introduce dependence of $\mu_{xh}$ across $x$ MacEachern (1999) uses a Gaussian process. De Iorio, Müller, Rosner and MacEachern (2004) construct the ANOVA DDP as a joint probability model for dependent random measures. They
consider a family of unknown probability measures $F_x$ indexed by categorical factors $x$. For example, in a clinical trial, $F_x$ might be the random effects distribution for patients with categorical covariates $x$. Covariates might include treatment levels, etc. Dependence across $\{F_x\}$ is induced by assuming ANOVA models on $\mu_{xh}$ across $x$.

6. MODEL VALIDATION

An interesting use of nonparametric Bayesian inference arises in model validation. One way to validate a proposed parametric model is to consider a nonparametric extension and report appropriate summaries of a comparison of the parametric and nonparametric fits.

Carota and Parmigiani (1996) and Carota, Parmigiani and Polson (1996) discuss such approaches using DP extensions and point out the limitations of formalizing the comparison with a Bayes factor. Due to the discrete nature of the Dirichlet process RPM inference is driven by the number of duplicates in the data set. They suggest, among other approaches, to consider KL-divergence of prior to posterior on the random probability model. Conigliani, Castro and O’Hagan (2000) discuss a similar approach, using fractional Bayes factors to summarize the comparison.

Berger and Guglielmi (2001) take up the same theme, but replace the DP prior with a PT model. To center the PT model at a parametric model $f(x|\theta)$ they construct PT’s with mean measure $f(x|\theta)$. They fix the nested partition sequence and set the parameters $\alpha_x$ for the random probabilities such that the desired mean is achieved. Computation of Bayes factors for the model validation is greatly simplified by the availability of a closed form expression for the marginal distribution under such PT models:

$$m(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \cdot \prod_{j=2}^{n} \prod_{m=1}^{m^*(x_j)} \alpha_{em}(x_j)(\alpha_{em-1}(x_j) + \alpha_{em-1}'(x_j)),$$

The $\alpha_x$ are the Beta distribution parameters in the definition of the PT, as defined in Section 2.3. The indices $e_m(x_j) = e_1 \cdots e_m$ identify the partitioning subset $B_{e_1 \cdots e_m}$ of level $m$ that contains $x_j$, that is, $x_j \in B_{e}$, and $\alpha_x'$ are the parameters of the posterior PT, given the observations $(x_1, \ldots, x_{j-1})$. The upper bound $m^*(x_j)$ in the product is the smallest level $m$ such that no $x_i$, $i < j$, belongs to the same partitioning subset $B_{e_m}(x_j)$ as $x_j$ at level $m$. The $\alpha$ sequences depend on the parameter $\theta$. Evaluation of Bayes factors of the parametric model versus nonparametric extension requires one more step of marginalization to marginalize w.r.t. $\theta$.

A related approach is pursued in Mazzuchi, Soofi and Soyer (2000). They consider parametric models defined as maximum entropy models in a moment class. This includes the exponential, Gamma, Weibull, normal, etc. By considering the posterior expected Kullback–Leibler divergence between the parametric model and a nonparametric extension centered at that parametric model they define a diagnostic of fit. For the nonparametric extension they use a DP model centered at the maximum entropy parametric model.

7. CONCLUSION

We have reviewed some important aspects of nonparametric Bayesian inference. Rather than attempt a complete catalog of existing methods we focused on typical modeling strategies in important inference problems. Also, we emphasized recent developments over a historical perspective. The chosen classification of Bayesian nonparametric approaches into the listed application areas is an arbitrary subjective choice, leading us to miss some interesting nonparametric Bayesian methods that did not fit cleanly into one of these arbitrary categories. Typical examples are Quintana (1998) and Lee and Berger (2001), discussing nonparametric approaches to modeling contingency tables and selection sampling, respectively.

An important aspect of nonparametric Bayesian inference that we excluded from the discussion are computational issues. Many approaches are driven by what are essentially computational concerns. Another important line of research that we excluded from the discussion are the many methods that are nonparametric in flavor even if they are not technically inference in infinite-dimensional parameter spaces. Typical examples are finite mixture models. Such models often provide flexible inference very much like corresponding nonparametric extensions.

Finally, we did not discuss methods that are nonparametric Baysian in the literal sense, rather than in the sense of the technical definition we gave in the Introduction. A typical example is Lavine (1995), who discusses inference based on a partial likelihood argument.
ACKNOWLEDGMENT

The first author was supported by NIH/HCI under Grand NIH R01CA75981. The second author supported in part by Grant FONDECYT 1020712.

REFERENCES


