THESIS DECLARATION

The undersigned

Wade, Sara Kathryn
PhD Registration Number: 1370113

Thesis Title: Bayesian Nonparametric Regression through Mixture Models

PhD in Statistics
24^{th} Cycle
Advisor: Professor Sonia Petrone
Year of Discussion: 2013

DECLARES

Under her responsibility:

1) that, according to the President's decree of 28.12.2000, No. 445, mendacious declarations, falsifying records and the use of false records are punishable under the penal code and special laws, should any of

these hypotheses prove true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree of 30th April 1999 protocol no. 224/1999, to keep copy of the thesis on deposit at the Biblioteche Nazionali Centrali di Roma e Firenze, where consultation is permitted, unless there is a temporary embargo in order to protect the rights of external bodies and industrial/commercial exploitation of the thesis;

3) that the Servizio Biblioteca Bocconi will file the thesis in its 'Archivio istituzionale ad accesso aperto' and will permit on-line consultation of the complete text (except in cases of a temporary embargo);

4) that in order keep the thesis on file at Biblioteca Bocconi, the University requires that the thesis be delivered by the candidate to Societá NORMADEC (acting on behalf of the University) by online procedure the contents of which must be unalterable and that NORMADEC will indicate in each footnote the following information:

    - thesis Bayesian Nonparametric Regression through Mixture Models;
    - by Sara Kathryn Wade;
    - discussed at Università Commerciale Luigi Bocconi - Milano in 2013;
    - the thesis is protected by the regulations governing copyright (law of 22 April 1941, no. 633 and successive modifications). The exception is the right of Universitá Commerciale Luigi Bocconi to reproduce the same for research and teaching purposes, quoting the source;

5) that the copy of the thesis deposited with NORMADEC by online procedure is identical to those handed in/sent to the Examiners and to any other copy deposited in the University offices on paper or

electronic copy and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (law of 22 April 1941, no. 633 and successive integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal and shall be exempt from any requests or claims from third parties;

7) that the PhD thesis is not the result of work included in the regulations governing industrial property, it was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results; it is not subject to patent or protection registrations, and therefore not subject to an embargo.

31 October, 2012

Wade, Sara Kathryn

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my gratitude to all who supported me and made this PhD thesis possible.

First of all, I must sincerely thank my advisor, Professor Sonia Petrone, for her invaluable advice. Her thoughtful motivations and careful analysis of problems invaluably enhanced this thesis and taught me to examine a problem from all angles. I am grateful for all the time she dedicated to working with me. As an advisor, she allowed me to feel comfortable asking even the most basic question and to form and voice my own opinions. In the end, she was more than an advisor to me; she not only continually helps to shape my career but, importantly, is a caring friend.

I am also very grateful for the excellent advice of my co-advisor, Professor Stephen G. Walker. His clever ideas are of great inspiration to me, and his entertaining company makes statistics and learning enjoyable.

My gratitude also extends to professors of Decision Sciences Department at Bocconi University for their wonderful courses and encouragement, including Professor Sandra Fortini, who carefully read through the thesis and provided very helpful and detailed comments.

Many thanks go to my PhD-mates, Silvia Mongelluzzo and Steffen Ventz, for their support, great company, and discussions on statistical problems and careers. I would also like to acknowledge the PhD students of other cycles for their friendship.

I must to express my gratitude to Isadora Antoniano Villalobos for her collaboration, discussions, hospitality, and friendship.

Thanks go to Dr. Giovanni Frisoni and Anna Caroli for the inspiration for the Alzheimer's disease studies.

I am grateful for the inspiration and suggestions of Professor David B. Dunson for Chapter 4 and helpful comments of the referees for Chapter 3.

Finally, I cannot forget my parents, who encouraged my love of mathematics and provided continual support, and Maurizio Alfonsi for all his support during these years.

Abstract

This thesis studies Bayesian nonparametric regression through mixture models. These types of models are highly flexible, yet also numerous, which raises the question of how to choose among the models for the application at hand. In answer to this question, we derive predictive equations for the conditional mean and density and carefully analyse the quantities involved. Our main contributions to the subject are a detailed study of the predictive performance of existing models, the identification of potential sources of improvement in prediction, and the development of novel procedures to improve prediction. The models developed are applied in three studies of Alzheimer's disease, with the aim of diagnosis of the disease based on AD biomarkers and investigation into the dynamics of AD biomarkers with increasing age.

Chapter 1

Introduction

This thesis is about Bayesian nonparametric regression models based on countable mixtures with an emphasis on examining the predictive performance of these models. The work is motivated from both a methodological and theoretical context and an applied problem concerning Alzheimer's disease.

1.1 Motivation

The linear regression model assumes the response variable Y is related to covariates x through a linear function with additive normal errors. It is the standard tool used in regression settings due to its simplicity, ease of interpretation, straightforward computations, and desirable asymptotic properties. However, in many situations, the assumptions of the standard linear regression model are unreasonable, leading to inadequate fitting of the data and poor predictive inference.

To relax the linearity assumption, a flexible approach consists in representing the regression function as a linear combination of basis functions. Indeed, most standard nonparametric methods, such as splines, wavelets, neural networks, and regression trees, can be represented in this fashion. Such methods can potentially approximate a wide range of regression

functions.

The literature on these types of models for curve or surface fitting, is huge; we mention some classical and Bayesian references for the interested reader. For classical splines models, we refer the reader to Wahba [1990], Hastie and Tibshirani [1990], and Friedman [1991]. Bayesian extensions of spline models can be found in a series of papers by Denison, Holmes, Mallick, and Smith, which are summarized in their book (Denison et al. [2002]), and by DiMatteo et al. [2001]. For a detailed reference of wavelets from both a classical and Bayesian perspective see Vidakovic [2009]. A nice discussion of neural networks with emphasis on Bayesian methods is given by Neal [1996], and a closely related frequentist method to neural networks is the projection pursuit regression of Friedman and Stuetzle [1981]. Breiman et al. [1984] is a standard reference for regression trees and a recent Bayesian extension can be found in Chipman et al. [2010]. In classical literature, two important estimators are obtained via kernel regression and local parametric regression (see Scott [1992], Chapter 8). In their book, Denison et al. [2002] also discuss Bayesian methods for local parametric regression. In Bayesian literature, another customary practice, that has gained recent attention, is to place a Gaussian process prior on the unknown regression function (see Rasmussen and Williams [2006]).

Yet, these various approaches are also limited in the sense that they only allow for flexibility in the mean. Many datasets also present departures from classical models such as non normality or multi-modality of the errors, or different variances, degrees of skewness, or tail behavior in different regions of the covariate space. To capture such behavior, a flexible approach for modeling the conditional density that allows both the mean and error distribution to evolve flexibly with the covariates is required.

For independent and identically distributed data, mixture models are an extremely useful tool for flexible density estimation due to their ability to approximate a large class of densities and their attractive balance between smoothness and flexibility in modeling local features. The form

of mixture model is given by

$$f_P(y) = \int K(y;\theta)dP(\theta), \tag{1.1}$$

where P is a probability measure on the parameter space Θ, $\mathcal{Y}$ is the sample space, and $K(y;\theta)$ is a kernel on $\mathcal{Y} \times \Theta$. The kernel, $K(y;\theta)$, is defined by

1) $\forall\, \theta \in \Theta$, $K(\cdot;\theta)$ is a density on $\mathcal{Y}$ with respect to the Lebesgue measure and

2) $K(y;\theta)$ is a measurable function of θ, where Θ is assumed to be a complete and separable metric space and equipped with its Borel σ-algebra.

In a Bayesian setting, this model is completed with a prior distribution on the mixing measure P. We will use the notation $\mathcal{M}(\Theta)$ to denote the set of probability measures on Θ and $\mathbf{P}$ to denote the random mixing measure taking values in $\mathcal{M}(\Theta)$. A common prior choice takes $\mathbf{P}$ as a discrete random measure with probability one. In this case, $\mathbf{P}$ has the following representation almost surely (a.s.)

$$\mathbf{P} = \sum_{j=1}^{J} w_j \delta_{\tilde{\theta}_j},$$

for some random atoms $\tilde{\theta}_j$ taking values in Θ and weights w_j such that $w_j \geq 0$ and $\sum_j w_j = 1$ (a.s.). The mixture model can then be expressed as a convex combination of kernels

$$f_P(y) = \sum_{j=1}^{J} w_j K(y;\tilde{\theta}_j). \tag{1.2}$$

Our interest is in extensions of this flexible class of models to address the problem of covariate-dependent density estimation. In this case, mixture models are not used to recover homogeneous sub-populations, but, rather, as a kernel method to obtain a flexible estimate of the covariate-dependent density. In general, the model may be extended in one of two

ways. The first approach is closely related to classical kernel regression methods and involves augmenting the observed data to include the covariates. The joint density is modelled by (1.2), i.e.

$$f_P(y,x) = \sum_{j=1}^{J} w_j K(y,x;\tilde{\theta}_j), \tag{1.3}$$

and conditional density estimates are obtained as a by-product of the joint density estimate through the equation

$$f_P(y|x) = \frac{\sum_{j=1}^{J} w_j K(y,x;\tilde{\theta}_j)}{\sum_{j'=1}^{J} w_{j'} K(x;\tilde{\theta}_{j'})}. \tag{1.4}$$

However, this approach unnecessarily requires the modelling of the marginal of X, when our interest is only on the conditional density. The second approach overcomes this by directly modelling the covariate-dependent density. In this case (1.1) is extended by allowing the mixing distribution to depend on x. Hence, for every $x \in \mathcal{X}$,

$$f_{P_x}(y|x) = \int K(y;x,\theta) dP_x(\theta).$$

Again, the Bayesian model is completed by assigning a prior distribution on the family $P_{\mathcal{X}} = \{P_x\}_{x\in\mathcal{X}}$ of covariate-dependent mixing probability measures. The notation $\mathbf{P}_{\mathcal{X}} = \{\mathbf{P}_x\}_{x\in\mathcal{X}}$ will be used to denote the family of random covariate-dependent mixing measures with realizations in $\mathcal{M}(\Theta)^{\mathcal{X}}$. If the prior gives probability one to the set of discrete probability measures, then (a.s.)

$$\mathbf{P}_x = \sum_{j=1}^{J} w_j(x)\delta_{\tilde{\theta}_j(x)},$$

and

$$f_{P_x}(y|x) = \sum_{j=1}^{J} w_j(x) K(y;x,\tilde{\theta}_j(x)), \tag{1.5}$$

where $\tilde{\theta}_j(x)$ takes values in Θ and the weights $w_j(x)$ are such that $w_j(x) \geq 0$ and $\sum_j w_j(x) = 1$ (a.s.) for all $x \in \mathcal{X}$.

Throughout the text, the first method (1.3) will be termed the *joint approach* and the second (1.5) will be called the *conditional approach*. Of course, the covariate may not be random. In this case, (1.5) is not a model for a conditional density but for a covariate-indexed density; thus, the phrase *conditional approach* is imprecise. Nevertheless, we will keep this terminology with this inconsistency in mind. Moreover, in order for (1.5) to define a proper random conditional density, $f_{P_x}(y|x)$, must be a measurable function of x almost surely. Assuming $\mathcal{X}$ is a complete and separable metric space, this condition is satisfied by defining $\mathbf{P}_{\mathcal{X}}$ to be measurable with respect to the Borel σ-algebra on $\mathcal{X}$ with probability one. Clearly, in the case when the covariate is non-random, the joint approach is not the natural choice, but it may still be used as a tool to obtain covariate-indexed density estimates.

The number of mixture components, J, in both the joint and conditional approach plays a key role in the flexibility of the model. Finite mixtures are defined with $J < \infty$ (see McLachlan and Peel [2000] for an overview). A recent reference for finite mixtures based on the joint approach is Norets and Pelenis [2012a], and references for finite mixtures based on the conditional approach, known as smooth mixtures of regressions, in econometrics literature, or mixture of experts, in machine learning literature, include Jacobs et al. [1991], Jacobs and Jordan [1994], and Geweke and Keane [2007]. For large enough J, (1.4) and (1.5) can both approximate a large class of covariate-dependent densities (Norets and Pelenis [2012a], Norets [2010]). However, they require either the choice of J, which in practice is chosen through post-processing techniques, or, in Bayesian setting, a prior on J, which requires posterior sampling of J.

Instead, nonparametric mixtures define $J = \infty$. The general models described by (1.3) and (1.5) with $J = \infty$ are the starting point for Bayesian nonparametric mixture models for regression, the focus of this thesis. The models are completed with a definition of the kernel and a prior choice for the weights and atoms. These types of models have become very popular

in Bayesian nonparametrics literature in the past decade, particularly after the introduction of Dependent Dirichlet processes (MacEachern [1999]). In Chapter 2, we provide an overview of the various proposals. The literature on this subject is rich, but it is somewhat fragmented; thus, Chapter 2 in itself provides a contribution to the subject by unifying existing literature.

Due to the large number of proposals, choosing among them for the application at hand can be a daunting task. Ideally, the chosen model should have good approximation properties to a large class of data-generating covariate-dependent densities and posterior consistency properties. Recently, these types of properties were explored for specific models based on the joint approach (Hannah et al. [2011]) and the conditional approach (Barrientos et al. [2012], Norets and Pelenis [2012b], Pati et al. [2012]). Posterior consistency is an interesting frequentist property that should be minimally satisfied, and we provide some discussion on the topic; however, it studies the behavior of the random conditional densities as the sample size goes to infinity. In practice, the sample size is finite, and a study of posterior consistency properties may hide what happens in the finite case.

This is a general theoretical issue, and it raises an important question: how do we choose among the different proposals of nonparametric models and priors from a *Bayesian* perspective? Although we do provide some discussion on frequentist asymptotic properties for the nonparametric models of interest, our main aim is to answer this question, and to do so, we adopt a natural approach from a Bayesian perspective that consists of a detailed study of properties based on *finite* samples. In particular, we carefully examine features of the model and prior and their effects on the predictive mean and density estimate for some new covariate values.

Our main contributions are 1) a detailed study of the predictive performance of existing models, 2) the identification of potential sources of improvement in prediction, and 3) the development of novel procedures to improve prediction. An interesting by-product of this research is the comparison of existing models including advantages and disadvantages depending on specific aspects of the observed data. In summary, we provide theoretical, methodological, and computational contributions that increase

the understanding of Bayesian nonparametric mixture models for regression and allow improved prediction.

1.2 Motivating application

The motivating application behind this work is to study Alzheimer's disease (AD) based on neuroimaging data. Alzheimer's disease is an irreversible, progressive brain disease that slowly destroys memory and thinking skills, and eventually even the ability to carry out the simplest tasks (ADEAR [2011]). It is a major public health concern, not only because of its damaging effects, but also because of its increasing prevalence and increasing life expectancy. In fact, in a study in 2007, Brookmeyer et al. estimated that over 26 million people worldwide were living with AD, and that number is predicted to grow to over 100 million by 2050.

To combat the disease, disease-modifying drugs or therapies are in great need. Drugs or therapies tend to be most effective in the early to mild stages of AD. Thus, early and differential diagnosis is also of great importance.

Unfortunately, definite diagnosis requires histopathologic examination of brain tissue, an invasive procedure typically only performed at autopsy. In practice, clinical diagnosis is based on a patient's history and symptoms, behavioral and cognitive tests, and visual examination of neuroimages, if available. The National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria, which is based on clinical and neuropsychological examination, can improve accuracy, but is time consuming. Several studies have followed patients to autopsy to estimate the accuracy of NINCDS-ADRDA criteria; the average sensitivity of the NINCDS-ADRDA criteria is 81% and the average specificity of the NINCDS-ADRDA criteria is 70% for the diagnosis of probable AD (Knopman et al. [2001]).

Alzheimer's disease is associated with the abnormal accumulation of the proteins amyloid-β (Aβ) and hyperphosphorylated tau (tau) leading

to impairment and loss of cognitive function, death of brain cells, and brain shrinkage. This neurobiological damage occurs gradually over time and is believed to start at early stages of the disease before the onset of clinical symptoms. In fact, some changes are believed to start possibly 20 years before the appearance of memory disturbances. Neuroimages are non-invasive tools that can be used to assess these changes and aid in diagnosis of the disease.

The first studies to examine the diagnostic ability of neuroimages focused on biomarkers based on structural Magnetic Resonance Images (sMRI). These biomarkers measure the volume or cortical thickness of specific brain structures and are computed based on automated or semi-automated approaches. Once these measures have been computed, studies use parametric methods, such as linear discriminant analysis or logistic regression, to estimate diagnostic accuracy.

However, the brain tissue loss associated with AD may occur only in part of the specified brain structure or may span multiple brain structures. Moreover, other types of neuroimaging, such as functional, microstructural, and amyloid imaging have recently been shown to be useful for diagnosis (Caroli and Frisoni [2009]). Thus, to improve diagnosis accuracy, there is a need to investigate the use of the entire sMRI and to combine this with data from other imaging techniques.

Clearly, incorporating the entire image as well as data from other imaging techniques will render the data increasingly complex and high-dimensional. In this setting, flexible nonparametric regression techniques are needed to capture complex interaction terms and encourage sparsity and dimension reduction. Furthermore, prior information about the relationship between disease status and its effects on the brain leads naturally to a Bayesian approach.

Neuroimages can also be of great use in clinical trials for AD; biomarkers based on neuroimaging data can be used as outcome measures to monitor disease progression, as inclusion criteria, and as disease-staging tools. Furthermore, they may better suited than clinical measures for disease staging and monitoring disease progression because of possible higher sen-

sitivity to changes due to drugs or therapies over shorter periods of time.

In order for biomarkers based on neuroimaging or biological data to be useful in clinical trials, their evolution over time needs to be well understood; those which change earliest and fastest should be used as inclusion criteria, those which change the most in the disease stage of interest should be used for disease monitoring, and all should be combined to assess the disease stage of the individual.

In a recent paper (Jack et al. [2010]), proposed a theoretical model for the evolution of the five most widely studied and well validated biomarkers. Their model assumed that biomarkers become abnormal in a time ordered manner with a sigmoidal path that varies in steepness across biomarkers. Frisoni et al. [2010] discussed the model in more detail, focusing on the evolution of biomarkers based on sMRI. They hypothesised a heterogeneous pattern for evolution across brain structures, with tissue loss first occurring in the entorhinal cortex, followed by the hippocampus, the temporal neocortex, and lastly, the whole brain. These structures are also hypothesized to display different sigmoidal shapes, with whole brain volume displaying the most gradual change over time.

Some recent studies have supported this model. Caroli and Frisoni [2010] and Sabuncu et al. [2011] assessed the fit of parametric sigmoidal curves, and Jack et al. [2012] considered a more flexible model based on additive cubic splines with three chosen knot points. However, even though the later approach is more flexible than the previous methods, there are still significant restrictions.

Flexible nonparametric regression techniques are needed in this setting to validate the proposed model and discover the nature of the hypothesized sigmoidal curve. Also, the model must be able to accommodate an evolving error distribution, which is likely in this situation due to the unobserved nature of the disease and additional factors, such as undiscovered neuroprotective genes. Furthermore, Bayesian methods can be used to incorporate prior information regarding the dynamics of the biomarkers.

In this work, we apply Bayesian nonparametric methods to study both the diagnosis of the disease and the dynamics of AD within the brain. In

particular, we consider Bayesian nonparametric mixture models of type (1.3) and (1.5) and focus on biomarkers based on sMRI. By relaxing the classic parametric assumptions that are typically assumed in literature, we are able to provide strong statistical support for existing theory and results as well as novel insight into the diagnosis and dynamics of the disease.

1.3 ADNI data

The data used for the Alzheimer's disease studies was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database which is publicly accessible at UCLA's Laboratory of Neuroimaging[1] .

The ADNI database contains neuroimaging, biological, and clinical data for AD, mild cognitive impairment (MCI), and cognitively normal (CN) patients. Summaries of neuroimages are also included, such as the volume and cortical thickness of various brain structures. The diagnosis and inclusion of the patients is based on a combination of NINCDS-

[1]The ADNI was launched in 2003 by the National Institute on Ageing (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $ 60 million, 5-year public- private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see `www.adni-info.org`.

ADRDA criteria and other clinical and neuropsychological tests, including the clinical dementia rating scale (CDR), the Wechsler memory scale (WMS), and the mini-mental state examination (MMSE). For more information, see `http://adni.loni.ucla.edu/wp-content/uploads/2010/09/ADNI_GeneralProceduresManual.pdf`.

1.4 Outline of thesis

Bayesian nonparametric mixture models for regression are the focus of this thesis, and we begin with a thorough review of models of this type in Chapter 2, providing a unifying framework for the models of interest. As this chapter clearly shows, the number of proposals and model choices is large and varied. Thus, to decide among the various choices in practice, a detailed understanding of properties of these models is needed.

In this direction, the next chapters carefully examine the predictive performance of these models. In particular, the prediction of models based on the joint approach (1.3) is studied in Chapter 4 and is further discussed in Chapter 5, along with a general discussion on the prediction of models based on the conditional approach (1.5). Then, in Chapter 6, we provide a closer examination of the prediction of models based on the conditional approach with flexible weight functions. Chapter 3, on the other hand, has a theoretical focus, but its developments are used to propose a novel model based on the joint approach with improved predictive performance in Chapter 4.

Chapters 3-6 contain the main contributions of this thesis, which are based on Wade et al. [2011], Wade et al. [2012], and Antoniano Villalobos et al. [2012], and an additional article which is joint work with Sonia Petrone and will be submitted shortly. We provide a brief summary of each chapter's contents.

In Bayesian nonparametric mixture models, the Dirichlet process (DP) is often used as a prior for a multivariate random probability measure. In Chapter 3, we discuss the rigidity of the DP in this case and propose an *enrichment* of the DP by extending the notion of enriched conjugate priors

to a nonparametric setting. The proposed process, the Enriched Dirichlet process (EDP), is more flexible, but is shown to maintain many desirables properties of the DP.

This process is then applied to a regression setting in Chapter 4. The chapter begins with a detailed examination of the predictive performance of Dirichlet process mixture models for the joint density of Y and X, with particular focus on the effect of increasing the dimension of X. We highlight some understated issues and to overcome them, propose to replace the DP with the EDP. We show the advantages of doing so through both predictive equations and two illustrative examples, a simulated example and a study into the diagnosis of AD based on a large number neuroimaging summaries.

In Chapter 5, an overlooked issue present in nonparametric mixture models, the huge dimension of the partition space, is underlined, and its effects on prediction are carefully studied through computations and illustrations. The predictive study also leads to interesting conclusions for the comparison of constant and covariate-dependent weights. We propose a novel covariate-dependent random partition model that reduces the size of the partition space and show that it maintains certain properties of random partition model implied by the DP. Advantages are demonstrated through simulated examples, and an application to examine the relationship between AD and the asymmetry of the hippocampus is presented.

Chapter 6 discusses models based on the conditional approach with covariate-dependent weights. The defined form of the covariate-dependent weight has important implications for prediction. We discuss limitations of current proposals and construct natural and interpretable weights based on normalization. A novel algorithm that deals with the normalizing constant is discussed in detail. Finally, two simulated examples and an interesting application to study the evolution of hippocampal volume as a function of age, sex, and disease status are presented.

Finally, Chapter 7 provides a final discussion and directions for future research.

Chapter 2

Review

Bayesian nonparametric mixture models for regression have gained much attention over the past decade. This chapter is dedicated to providing a review of the literature and unifying framework for the various proposals.

2.1 Dirichlet process

We begin with a review of the Dirichlet process (DP) because it is commonly used in many of the models of interest. The Dirichlet process was first introduced by Ferguson [1973] and is now the most popular prior in Bayesian nonparametrics. For a complete and separable metric space Θ, the DP defines a distribution on $\mathcal{M}(\Theta)$, the space of probability measures on Θ, and its Borel σ-algebra under weak convergence. It is characterized by the fact that the finite dimensional distributions of the probability over any measurable partition are Dirichlet, with consistent parameters. In more detail, a random probability measure $\mathbf{P}$ on Θ is a Dirichlet process with parameters $\alpha > 0$ and $P_0 \in \mathcal{M}(\Theta)$, denoted by $\mathrm{DP}(\alpha P_0)$, if for any finite measurable partition $(C_1, \ldots, C_m)$ of Θ,

$$(\mathbf{P}(C_1), \ldots, \mathbf{P}(C_m)) \sim \mathrm{Dir}(\alpha P_0(C_1), \ldots, \alpha P_0(C_m)).$$

The Dirichlet process has many desirable properties including easy elic-

itation of its parameters, large support, and conjugacy. Another important property that is frequently utilized is the almost sure discrete nature of $\mathbf{P}$. In fact, Sethuraman (1994) showed that the DP can also be characterized through the stick-breaking representation

$$\mathbf{P} = \sum_{j=1}^{\infty} w_j \delta_{\tilde{\theta}_j},$$

where

$$w_1 = v_1,$$

$$w_j = v_j \prod_{j'<j} (1 - v_{j'}) \quad \text{for } j > 1,$$

$$v_j \overset{iid}{\sim} \text{Beta}(1, \alpha),$$

and independent of (v_j),

$$\tilde{\theta}_j \overset{iid}{\sim} P_0.$$

We should comment that, here and throughout the rest of the text, we use, with a slight abuse of notation, $\theta \sim P$ to mean that θ is distributed according to the distribution function associated to the probability measure P. The term stick-breaking is used because this construction of the weights can be visualized through sequential breaks of a stick of length one. In particular, the first weight is the length of the first broken piece of the stick, the second is the length of a break of the remaining stick, etc., where v_j represents the proportion of the break at step j. More general stick-breaking constructions are reviewed and given in Ishwaran and James [2001].

Assuming $\theta_i | P \overset{iid}{\sim} P$ and $\mathbf{P} \sim \text{DP}(\alpha P_0)$, since $\mathbf{P}$ is discrete with probability one, it implies positive probabilities of ties among the sample $(\theta_1, \ldots, \theta_n)$. Let k_n then denote the number of unique values among the observations and $(\theta_1^*, \ldots, \theta_{k_n}^*)$ denote the unique values. The predictive distribution of the observations is given by the Pólya urn scheme

(Blackwell and MacQueen [1973]),

$$\theta_1 \sim P_0,$$
$$\theta_{n+1} \mid \theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha+n} P_0 + \sum_{j=1}^{k_n} \frac{n_{n,j}}{\alpha+n} \delta_{\theta_j^*},$$

where $n_{n,j} = \sum_{i=1}^{n} \mathbf{1}(\theta_i = \theta_j^*)$, is the number of observations that are equal to the j^{th} unique value. For ease of notation, we drop the subscript n from $(k_n, n_{n,j})$ when the sample size is understood. The observations $(\theta_1, \ldots, \theta_n)$ can be equivalently parametrized in terms of the independent vectors $(s_1, \ldots, s_n)$ and $(\theta_1^*, \ldots, \theta_k^*)$, where

$$\mathbf{s}_1 \sim \delta_1, \tag{2.1}$$
$$\mathbf{s}_{n+1} \mid s_1, \ldots, s_n \sim \frac{\alpha}{\alpha+n} \delta_{k+1} + \sum_{j=1}^{k} \frac{n_j}{\alpha+n} \delta_j, \tag{2.2}$$
$$\theta_j^* \overset{iid}{\sim} P_0 \quad \text{for } j = 1, \ldots k,$$

and $\theta_i = \theta_j^*$ if $s_i = j$. An entertaining interpretation of the distribution of $(s_1, \ldots, s_n)$ described by (2.1) and (2.2) is given by the Chinese restaurant process (see Pitman [1995] for more details). Subjects sequentially enter a Chinese restaurant, where the first subject sits at the first table. The second subject will sit at the first table with probability proportionally to 1 or at a new table with probability proportional to α. This process is repeated, so that, if, after n subjects, k tables are occupied with $n_1, \ldots, n_k$ subjects at each table, the $n+1^{th}$ subject will sit at the j^{th} occupied table with probability proportional to n_j or at a new table with probability proportional to α.

Random partition models define the distribution of the partition of n subjects into k clusters (see Quintana [2006]). The DP implicitly defines a random partition model, through the joint distribution of $(s_1, \ldots, s_n) = \rho_n$. From (2.1) and (2.2), we have that

$$p(\rho_n) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \alpha^k \prod_{j=1}^{k} \Gamma(n_j).$$

In Bayesian nonparametric mixture models, the Dirichlet process is commonly chosen as the prior for the mixing measure. This type of model was first introduced and studied by Lo [1984]. In this case, we observe $(y_1, \dots y_n)$, and $(\theta_1, \dots, \theta_n)$ represent the latent subject-specific parameters, where we assume

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} F(\cdot|\theta_i), \\ \theta_i|P &\overset{iid}{\sim} P, \\ \mathbf{P} &\sim \mathrm{DP}(\alpha P_0). \end{aligned}$$

Integrating out the $(\theta_1, \dots, \theta_n)$, we have that given P, the Y_i are independent with density

$$f_P(y) = \int_\Theta K(y;\theta) dP(\theta) = \sum_{j=1}^{\infty} w_j K(y;\tilde{\theta}_j), \tag{2.3}$$

where $K(\cdot;\theta)$ is the density of $F(\cdot|\theta)$.

The DP mixture model (2.3) for density estimation is very flexible, and the stick-breaking construction, Pólya urn scheme, and random partition model defined by the DP are important in computations. As we will see in the next sections, these representations are also frequently extended to define proposals of Bayesian nonparametric mixture models for covariate-dependent density estimation.

2.2 Joint approach

A simple extension of DP mixture models for density estimation to covariate-dependent density estimation augments the observations to include the covariates. The joint density of Y and X is modelled flexibly through

$$f_P(y,x) = \sum_{j=1}^{\infty} w_j K(y,x;\tilde{\theta}_j), \tag{2.4}$$

where P is a realization of the random probability measure $\mathbf{P}$. Most proposals use a DP as the prior of $\mathbf{P}$, but more generally, $\mathbf{P}$ may be

defined as

$$\mathbf{P} = \sum_{j=1}^{\infty} w_j \delta_{\tilde{\theta}_j},$$

for some weights such that $w_j > 0$ and $\sum_{j=1}^{\infty} w_j = 1$ (a.s) and atoms defined, independently of (w_j), by $\theta_j \overset{iid}{\sim} P_0$.

Inference is carried out as for the joint density, and conditional density estimates are obtained from the posterior inference based on the joint model. In particular, the model for the conditional density is

$$f_P(y|x) = \frac{\sum_{j=1}^{\infty} w_j K(y, x; \tilde{\theta}_j)}{\sum_{j'=1}^{\infty} w_{j'} K(x; \tilde{\theta}_{j'})}.$$

The multivariate density $K(y, x; \theta)$ can be expressed as the product of the marginal density on $\mathcal{X}$ and the conditional density on $\mathcal{Y}$ given x and, in most cases, reparametrized so that the marginal and conditional density each depend on their own parameter θ_x and $\theta_{y|x}$, respectively. To simplify notation, throughout the rest of the text, the parameter θ_x will be denoted by ψ, with the marginal density on $\mathcal{X}$ denoted by $K(x; \psi)$, and the parameter $\theta_{y|x}$ will be denoted simply by θ, with the conditional density denoted by $K(y; x, \theta)$. In this case, the model for the conditional density can be equivalently written as

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y; x, \tilde{\theta}_j),$$

where

$$w_j(x) = \frac{w_j K(x; \tilde{\psi}_j)}{\sum_{j'=1}^{\infty} w_{j'} K(x; \tilde{\psi}_{j'})},$$

and

$$\mathbf{P}_x = \sum_{j=1}^{\infty} w_j(x) \delta_{\tilde{\theta}_j}.$$

Thus, (2.4) implicitly defines a model for the conditional density of form specified in the conditional approach.

This approach was first introduced by Müller et al. [1996], who assume a multivariate normal kernel within component for a continuous response

and continuous covariates and use a DP prior for $\mathbf{P}$. In recent literature, extensions and further discussions of this model have received increasing attention; most employ a DP prior and discuss alternative kernel choices or examine properties. Shahbaba and Neal [2009] and Hannah et al. [2011] discuss extensions for other types of responses through different kernel functions. Kang and Ghosal [2009] employ some frequentist techniques in estimation and discuss advantages over other flexible approaches for multivariate covariates, such as multivariate splines, that rely on partitioning the covariate space. Park and Dunson [2010] and Müller and Quintana [2010] examine the covariate-dependent urn scheme implicitly defined by the model. A nice application of the model to study the relationship between water quality and pregnancy outcomes is given in Dunson and Herring [2006]. Taddy and Kottas [2010] use the model to study quantile regression. In Bhattacharya et al. [2012], an alternative kernel to achieve dimension reduction of x is explored. An alternative prior choice for the mixing measure, the skewed Dirichlet process (Iglesias and Quintana [2009]), is discussed in Quintana [2011].

2.2.1 Consistency

Frequentist properties, such as posterior consistency, provide validation for the models of interest. For the joint approach, as a first step, one may interested in posterior consistency of the joint density. Posterior consistency of DP mixture models for univariate density estimation is studied in Ghosal et al. [1999], Ghosal and van der Vaart [2001], Ghosal and van der Vaart [2007], Tokdar [2006], and Walker et al. [2007]. Results for multivariate density estimation appeared later in Wu and Ghosal [2008], Wu and Ghosal [2010], and Tokdar [2011]. In these studies, one assumes that given f_P, the observations $Z_i = (Y_i, X_i)$ are i.i.d with density

$$f_p(z) = \int K(z;\theta) dP(\theta),$$

and $\mathbf{P} \sim \mathrm{DP}(\alpha P_0)$. If, in reality, the data are independently generated from some density f_0, one is interested in what kind of conditions on the

data-generating density f_0; on the kernel $K(\cdot;\theta)$; and on the parameters of the DP (α, P_0) imply that the posterior of the random density concentrates around the true density with a high probability as the sample size goes to infinity, or more formally,

$$Q_f(U_\epsilon(f_0)|Z_{1:n}) \to 1 \quad \text{a.s. } P_{f_0}^\infty,$$

for any $\epsilon > 0$, where Q_f denotes the law of random density of the DP mixture model; P_{f_0} denotes the probability measure associated to f_0; and $U_\epsilon(f_0)$ denotes a neighborhood of f_0 of size ϵ. For weak consistency, the neighborhood $U_\epsilon(f_0)$ is defined by

$$U_\epsilon(f_0) = \left\{ f \in \mathcal{F} : \mid \int g_i(z) f(z) dz - \int g_i(z) f_0(z) dz \mid < \epsilon, \quad i = 1, \ldots, m \right\},$$

where $\mathcal{F}$ is the set of densities on $\mathcal{Y}$ (with respect to the Lebesgue measure) and $g_i(\cdot)$ are bounded, continuous functions on $\mathcal{Y}$. For strong consistency, the neighborhood is often defined by

$$U_\epsilon(f_0) = \left\{ f \in \mathcal{F} : \int \mid f(z) - f_0(z) \mid dz < \epsilon \right\}.$$

For the joint DP mixture model, Hannah et al. [2011] prove weak consistency of the joint density for certain kernel choices and under specific conditions on f_0. Next, they show that posterior consistency of the joint density has implications for the regression function. In particular, with some additional mild conditions, it follows that the estimated regression function converges pointwise to $\mathrm{E}_{P_{f_0}}[y|x]$, i.e.

$$\mathrm{E}[Y|x, Y_{1:n}, X_{1:n}] \to \mathrm{E}_{P_{f_0}}[Y|x] \quad \text{a.s. } P_{f_0}^\infty,$$

where $\mathrm{E}[Y|x, Y_{1:n}, X_{1:n}]$ is the prediction under the DP mixture model.

2.3 Conditional approach

We are only interested in the conditional density, and, in this case, modelling also the marginal density of X is an unnecessary complication. The

conditional approach overcomes this by directly modelling the collection of conditional densities $\{f(y|x)\}_{x\in\mathcal{X}}$. For this approach, classic nonparametric mixtures for density estimation can be extended to define a flexible model for $\{f(y|x)\}_{x\in\mathcal{X}}$ by allowing the mixing measure to depend on x:

$$f_{P_x}(y|x) = \int K(y;x,\theta)dP_x(\theta). \tag{2.5}$$

The task is then to give a prior on $\{\mathbf{P}_x\}_{x\in\mathcal{X}}$ so that the random probability measures are dependent across x. The covariate-dependent random probability measures are assumed to be discrete (a.s), and thus, they have the following representation

$$\mathbf{P}_x = \sum_{j=1}^{\infty} w_j(x)\delta_{\tilde{\theta}_j(x)}. \tag{2.6}$$

By introducing dependency in the weights and the atoms, it is possible to obtain inference without the requirement of repeated observations at each covariate value.

Some early proposals are closely related to (2.5) and (2.6), but the general model was introduced by MacEachern in 1999 and 2000. Since then, the subject has become increasingly popular. Existence of the family of random probability measures in (2.6) was discussed by MacEachern [2000], and relies on the existence of the collection of stochastic processes $(w_j(\cdot), \theta_j(\cdot))$. Most proposals fall into one of two important subclasses: 1) models with flexible covariate-dependent atoms but simple weights and 2) models with flexible covariate-dependent weights and simple atoms.

2.3.1 Early proposals

A first proposal to define a prior for the collection of random probability measures $\{\mathbf{P}_x\}_{x\in\mathcal{X}}$ was given by Cifarelli and Regazzini [1978], where the focus was on discrete covariates. They introduced dependence between a vector of random probability measures through the base measure of a Dirichlet process. Their proposal extends Antoniak's (1974) mixture of

Dirichlet processes. In particular, assuming $\mathcal{X} = \{1, \ldots, M\}$ for some finite M, the law of the M-vector of random probability measures is

$$\mathbf{P}_1, \ldots, \mathbf{P}_M | u_1, \ldots, u_M \sim \prod_{x=1}^{M} \mathrm{DP}(\alpha(u_x, \cdot)), \tag{2.7}$$

where

$$u_1, \ldots, u_M \sim H,$$

for some distribution H. Typically, $\alpha(u_x, \cdot)$ is assumed to have the form $\alpha_x P_0(\cdot|u_x)$. In terms of equation (2.5), this implies that the weights are allowed to vary with x, but are constructed independently across x, in accordance with the DP. Thus, dependence is induced through the covariate dependent atoms, where

$$\tilde{\theta}_j(x)|u_x \overset{ind}{\sim} P_0(\cdot|u_x).$$

This idea was applied in regression and ANOVA settings by Cifarelli et al. [1981], for studying the search of the optimal dose in Muliere and Petrone [1993], and to address change point problems in Mira and Petrone [1996]. In this approach, since the weights are independent across x, multiple observations at each covariate value are needed for inference. For example, in Muliere and Petrone [1993], only a finite number of doses x were possible, and they assume $u_x = \beta \quad \forall\, x \in \mathcal{X}$ and

$$\tilde{\theta}_j(x)|\beta \overset{ind}{\sim} N(\underline{X}\beta, \sigma^2),$$

where $\underline{X} = (1, x')$.

However, in these studies, the idea was to use (2.7) to directly define a model for the collection of conditional distribution functions, not through a mixture. A limitation of this approach is that the nature of the dependence is restricted to the form specified in the base measure, and a deeper discussion of drawbacks of this approach is given in Petrone and Raftery [1997].

An early proposal for a mixture model of type (2.5) defines the weights as constant functions of x and assumes that the kernel $K(y; x, \theta(x))$ is the

standard linear regression model. In this case, the model corresponds to an infinite mixture of linear regression models. One can imagine a non-homogeneous population, where a subject's response behaviour may be described by one of the models in the infinite collection of linear regression models, and allocation to a specific component is independent of x. More formally, the model is

$$f_P(y|x) = \sum_{j=1}^{\infty} w_j \mathrm{N}(y; \underline{X}\tilde{\beta}_j, \tilde{\sigma}_j^2),$$

where P denotes a realization of $\mathbf{P}$ and

$$\mathbf{P} = \sum_{j=1}^{\infty} w_j \delta_{(\tilde{\beta}_j, \tilde{\sigma}_j^2)}.$$

The notation $\mathrm{N}(\cdot; \mu, \sigma^2)$ denotes the normal density with a mean of μ and variance of σ^2. The typical choice for the law of $\mathbf{P}$ is the Dirichlet process. An early overview of Dirichlet process mixtures of linear models, with applications, is the article by West et al. [1994].

2.3.2 General model

In MacEachern [1999] and in a more detailed technical report, MacEachern [2000], the general and flexible model (2.5) was introduced. MacEachern was specifically interested in models that assumed the marginal of P_x is a Dirichlet process, which was chosen because of the desirable properties discussed in Section 2.1 as well as the availability of computational procedures for inference.

MacEachern's general class of Dependent Dirichlet process (DDP) assume that each $w_j(\cdot)$, for $j = 1, 2, \ldots$, is a stochastic process on $\mathcal{X}$ with the stick-breaking construction

$$w_1(x) = v_1(x),$$

$$w_j(x) = v_j(x) \prod_{j'<j} (1 - v_{j'}(x)) \quad \text{for } j > 1,$$

where each $v_j(\cdot)$ is a stochastic process on $\mathcal{X}$ with marginal distributions

$$v_j(x) \sim \text{Beta}(1, \alpha(x)),$$

and the $v_j(\cdot)$ are independent across j. The atoms, $(\tilde{\theta}_j(\cdot))$, are independent across j, and for each j, $\tilde{\theta}_j(\cdot)$ is a stochastic process on $\mathcal{X}$ with marginals P_{0x} for $x \in \mathcal{X}$. Additionally, the atoms $(\tilde{\theta}_j(\cdot))$ are independent of $(v_j(\cdot))$.

Applications of the fully flexible DDP model, or more generally, models with fully flexibly weights and atoms, are hard to find. One example is the model for spatial applications proposed by Duan et al. [2007]. This lack of proposals for fully flexibly models is due to interpretability issues, computational complexities, and the fact that desirable theoretical properties are still available with simpler constructions.

In fact, Barrientos et al. [2012] show full weak support of the random covariate-dependent mixing measures $\{\mathbf{P}_x\}_{x\in\mathcal{X}}$ for the general DDP model and also for two simplified versions which assume constant weights or constant atoms; that is, recalling that $\mathcal{M}(\Theta)$ is the set of probability measures on Θ, the topological support, assuming the product Borel σ-algebra under weak convergence, is $\mathcal{M}(\Theta)^{\mathcal{X}}$ (assuming, of course, that the topological support of P_{0x} is Θ for all x). In any case, only reasonable conditions on the stochastic process $v_j(\cdot)$ and $\tilde{\theta}_j(\cdot)$ are required. Moreover, for the general DDP mixture model, as well as for the two simplified DDP mixture models, they also demonstrate that a large class of data-generating conditional densities is contained in the support of the random conditional densities $\{f_{\mathbf{P}_x}(\cdot|x)\}_{x\in\mathcal{X}}$, where, on $\mathcal{F}^{\mathcal{X}}$, the product space of densities on $\mathcal{Y}$, they consider neighborhoods defined by the product Hellinger metric and by the product Kullback-Leibler divergence. In this first case, some additional constraints on the basis kernel for y, $K(y; x, \theta)$, are required, and for the second, stronger constraints on $K(y; x, \theta)$ are needed.

2.3.3 Covariate-dependent atoms

An important simplified class of models assumes flexible covariate-dependent atoms but constant weights:

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j K(y; x, \tilde{\theta}_j(x)), \tag{2.8}$$

where P_x is a realization of

$$\mathbf{P}_x = \sum_{j=1}^{\infty} w_j \delta_{\tilde{\theta}_j(x)}.$$

In most cases, $K(y; x, \theta(x))$ is defined so that the regression function $\mathrm{E}[y|x, P_x]$ is described by one of infinite collection of possible mean functions $\tilde{\theta}_j(x)$, with probability w_j. It is important to note that this probability of allocation to a specific mean function is independent of x. These models are attractive because inference can be carried out using any of the established algorithms for Bayesian nonparametric mixture models (see e.g. MacEachern [1994], Ishwaran and James [2001], Neal [2000], Papaspiliopoulos and Roberts [2008], Kalli et al. [2011]), resulting in much simpler computations.

An important example of (2.8) is the *single-p* DDP, which defines w_j in accordance with the DP. It is a special case of the DDP models introduced by MacEachern [1999] and the model he employed in applications. Single-p DDP mixtures are popular and have been successfully applied to address a wide range of problems from classical regression problems (MacEachern [2000], MacEachern [2001]) to ANOVA (De Iorio et al. [2004]), spatial modeling (Gelfand et al. [2005]), time series (Rodriguez and Horst [2008]), discriminant analysis (De La Cruz et al. [2007]), longitudinal analysis (Müller et al. [2005]), and survival analysis (De Iorio et al. [2009], Jara et al. [2010]).

For continuous covariates and a continuous response, the most popular single-p DDP model is

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j N(y; \tilde{\mu}_j(x), \tilde{\sigma}_j^2), \tag{2.9}$$

where $\tilde{\mu}_j(\cdot)$ are independent Gaussian processes with a mean function of $m(\cdot)$ and covariance function of $c(\cdot,\cdot)$, denoted by $\mathrm{GP}(m,c)$. Even in this simplified model (2.9), there are various choices for $m(\cdot)$ and $c(\cdot,\cdot)$. For example, MacEachern [2000] studies the log area of Romanesque churches given the log perimeter, and MacEachern [2001] studies biology exam scores given previous exam scores, where in both applications, he assumes a linear mean function of the Gaussian processes, i.e. $m(x) = \underline{X}\beta$, and an exponential variogram for the covariance function of the Gaussian processes, i.e.

$$c(x,x') = (c_0 - c_1)(1 - \exp(-\tau||x - x'||)) + c_1\mathbf{1}(||x - x'|| > 0).$$

Assuming that $m(\cdot)$ is a linear function expresses the belief that within each component, the regression function is close to linear with a Gaussian process residual. This model is also applied in Gelfand et al. [2005], where x represents the spatial location of an observation. In this example, the Gaussian processes are specified to have mean zero with a squared exponential covariance function,

$$c(x,x') = c\exp(-\tau||x - x'||^2).$$

De Iorio et al. [2004] focus on discrete covariates and show that in this setting, the single-p DDP is equivalent to a DP mixture of linear regression models under a transformation, $\phi(\cdot)$, of x into a higher-dimensional space. The general model for discrete covariates and a continuous response is

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j N(y; \tilde{\beta}_j'\phi(x), \tilde{\sigma}_j^2). \tag{2.10}$$

The most flexible choice of $\phi(\cdot)$ transforms the p-dimensional discrete vector x into a $M_1 * \ldots * M_p$-dimensional vector of zeros apart from a single element of one indicating the categories of the p covariates, where M_h is the number categories of the h^{th} covariate.

Extensions of (2.9) and (2.10) for other response types involve simply replacing the normal kernel with the appropriate kernel. For example, in

De Iorio et al. [2004], two datasets are considered; in the first, a multivariate binary response is present, and the second contains a functional continuous response that represents white blood cell count over time. The covariates are discrete, representing treatment type and the dose level of a drug. Thus, model (2.10) is employed and, in the first example, extended by replacing the local linear regression model $N(y; \tilde{\beta}_j'\phi(x), \tilde{\sigma}_j^2)$ with an ordered probit model. In the second, y is indexed by an additional variable t, representing time, and the model is extended by replacing the local mean $\tilde{\beta}_j'\phi(x)$ in (2.10) with some specified function of t and $\tilde{\beta}_j'\phi(x)$. A similar extension is discussed in De La Cruz et al. [2007], where the response represents the level of a specific hormone over time and x is a binary indicator for normal pregnancy.

In general, the procedure used in (2.10) of mapping x to a high-dimensional vector may also be used for continuous covariates by defining an appropriate transformation function. In fact, models that define the mean functions $\tilde{\mu}_j(x)$ through Gaussian processes (2.9) can be represented in terms of models with mean functions of the form in (2.10), $\tilde{\beta}_j'\phi(x)$, because $\tilde{\mu}_j(x)$ can be equivalently written as $\tilde{\beta}_j'\phi(x)$ where $\phi(x)$ transforms x into a possibly infinite dimensional space whose transformation is defined by the covariance function of the Gaussian process. More specifically, if $c(\cdot, \cdot)$ is the covariance function, then $c(x_1, x_2) = \phi(x_1)'\phi(x_2)$. See Section 4.3 of Rasmussen and Williams [2006] for examples.

To accommodate continuous and discrete covariates, an appropriate transformation needs to be defined. For example, in De Iorio et al. [2009], flexible mean functions for discrete covariates and linear mean functions for the continuous covariates are used, so that $\tilde{\mu}_j(x) = \tilde{\beta}_{d,j}'\phi(x_d) + \tilde{\beta}_{c,j}'x_c$, where x_d and x_c represent the discrete and continuous covariates, respectively. Instead, in Jara et al. [2010], they use linear mean functions for both the discrete and continuous covariates, i.e. $\tilde{\mu}_j(x) = \underline{X}\tilde{\beta}_j$. Both consider applications to survival analysis where the former studies the survival time for cancer patients given the dose level of a drug (discrete), estrogen receptor status (discrete), and tumor size (continuous), and the latter studies time to dental carry given information of dental hygiene (mostly

binary apart from the age at the start of brushing). Note that when the transformation is simply the identity function, i.e. $\phi(x) = x$, so that the mean functions are linear, the model is equivalent to the mixture of linear regression models discussed in Section 2.3.1. For increased model flexibility, higher-dimensional transformations are needed. De Iorio et al. [2009] mention including higher-order terms for the continuous covariates, and and Jara et al. [2010] comment that $\phi(x_c)$ may be defined through B-splines. For flexible interactions terms, an appropriate transformation is needed.

In most applications, the weights are defined through the DP, but this may also be extended. For example, Jara et al. [2010] examine the use of both the Dirichlet process and the two-parameter Poisson-Dirichlet process (Pitman and Yor [1997]). The latter assumes the usual stick-breaking construction for the weights,

$$w_1 = v_1,$$
$$w_j = v_j \prod_{j'<j} (1 - v_{j'}) \quad \text{for } j > 1,$$

where

$$v_j \overset{ind}{\sim} \text{Beta}(1 - a, b + ja),$$

for $0 \leq a < 1$ and $b > -a$.

Consistency

For conditional density estimation, the notion of posterior consistency requires one to imagine that the data are generating by a set of conditional densities $\{f_{0\,x}\}_{x \in \mathcal{X}} = f_{0\,\mathcal{X}}$; that is, the Y_i given x_i are generated independently from $f_{0\,x_i}$. Posterior consistency results in this setting are quite recent, and most rely on posterior consistency theorems formulated for joint densities. This requires the additional assumption that X_i are generated from some marginal density $h(x)$. It is important to note that the posterior of the conditional density does not involve $h(x)$; the data-generating marginal density $h(x)$ is only introduced as a tool for studying posterior consistency.

In this case, posterior consistency at the data-generating conditional densities $f_{0\,\mathcal{X}}$ requires that

$$Q_{f_{\mathcal{X}}}(U_\epsilon(f_{0\,\mathcal{X}})|Y_{1:n}, X_{1:n}) \to 1 \quad \text{a.s. } P_{f_0}^{\infty},$$

for any $\epsilon > 0$, where $Q_{f_{\mathcal{X}}}$ denotes the law of random conditional densities defined by the general model (2.5); $U_\epsilon(f_{0\,\mathcal{X}})$ denotes a neighborhood of $f_{0\,\mathcal{X}}$; and P_{f_0} denotes the probability measure associated to data-generating *joint* density $f_0(y, x) = f_{0\,x}(y|x)h(x)$. Again, one is interested in discovering the conditions on $f_{0\,\mathcal{X}}$; the kernel $K(y; x, \theta(x))$; and the random conditional probability measures $\mathbf{P}_{\mathcal{X}}$ that lead to posterior consistency. For weak consistency, the neighborhood $U_\epsilon(f_{0\,\mathcal{X}})$ is defined by

$$U_\epsilon(f_{0\,\mathcal{X}}) = \{f_{\mathcal{X}} \in \mathcal{F}_c^{\mathcal{X}} : | \int g_i(y, x) f(y|x) h(x) dy dx - \int g_i(y, x) f_0(y|x) h(x) dy dx \,| < \epsilon, \quad i = 1, \ldots, m\},$$

where $\mathcal{F}_c^{\mathcal{X}}$ is the set of conditional densities and $g_i(\cdot, \cdot)$ are bounded, continuous functions on $\mathcal{Y} \times \mathcal{X}$. For strong consistency, the neighborhood may be defined by

$$U_\epsilon(f_{0\,\mathcal{X}}) = \{f_{\mathcal{X}} \in \mathcal{F}_c^{\mathcal{X}} : \int \left(\int \mid f(y|x) - f_0(y|x) \mid dy \right) h(x) dx < \epsilon\}.$$

or as

$$U_\epsilon(f_{0\,\mathcal{X}}) = \{f_{\mathcal{X}} \in \mathcal{F}_c^{\mathcal{X}} : \sup_{x \in \mathcal{X}} \int \mid f(y|x) - f_0(y|x) \mid dy < \epsilon\}.$$

In a recent paper, Pati et al. [2012] demonstrate weak and strong consistency of models of type (2.9) for a general class of bounded data-generating densities satisfying certain tail conditions. For weak consistency, only continuity and approximation properties for $\tilde{\mu}_j(\cdot)$ are required with any set of weights that sum to one (a.s.). For strong consistency, more stringent conditions are required for both the mean functions and

the weights. The mean functions are carefully specified as

$$\begin{aligned}\tilde{\mu}_j(x) &= \underline{X}\tilde{\beta}_j + \tilde{\eta}_j(x),\\ \tilde{\eta}_j(x)|\tau &\overset{iid}{\sim} \mathrm{GP}(0, c),\\ c(x_1, x_2) &= c\exp(-\tau||x_1 - x_2||^2),\\ \tau^{p(1+\eta_2)/\eta_2} &\sim \mathrm{Gamma}(a, b),\end{aligned}$$

where p is the dimension of x and τ, η_2, a, b are fixed positive constants. Further conditions are also required on the priors of $\tilde{\beta}_j$ and $\tilde{\sigma}_j^2$. The weights must decay rapidly enough, and the usual DP weights do not actually satisfy their condition. The condition on the weights limits model complexity so that with flexible mean functions, only a few components will have relatively high weights.

To our knowledge, there are currently no results on posterior consistency of models that define the flexible mean functions through higher-order transformation functions of x (2.10) and thus, no results for models of type (2.8) when discrete covariates or both discrete and continuous covariates are present.

2.3.4 Covariate-dependent weights

Recent developments explore the idea of covariate-dependent weights. The general model (2.5) is usually simplified by assuming that the atoms do not to depend on the covariates,

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y; x, \tilde{\theta}_j), \tag{2.11}$$

where P_x is a realization of

$$\mathbf{P}_x = \sum_{j=1}^{\infty} w_j(x)\delta_{\tilde{\theta}_j}.$$

The idea behind these models is that the response distribution at x can be described by an infinite collection of parametric regression models, and

that the local parametric regression models used to describe the response distribution at x depend on the location of x in the covariate space.

The main constraint in this case is given by the need to specify a prior such that $\sum_j w_j(x) = 1$ for all $x \in \mathcal{X}$. In literature, the technique used to explicitly define $w_j(x)$ and satisfy this constraint is based on the stick-breaking representation:

$$w_1(x) = v_1(x),$$

$$w_j(x) = v_j(x) \prod_{j'<j} (1 - v_{j'}(x)) \quad \text{for } j > 1,$$

where $0 \leq v_j(x) \leq 1$ a.s. for all j and x. The various models present in literature differ in the definition of $v_j(x)$, and for each proposal, various model choices regarding hyperparameters and functional shapes are needed. Without loss of generality, we denote the additional parameters used to define $v_j(x)$ by the same symbol $\tilde{\psi}_j$ in all constructions.

One of the first approaches was developed by Griffin and Steele [2006], who incorporate dependency in the weights by reordering the v_j's based on x. One way to accomplish this is to associate each $(v_j, \tilde{\theta}_j)$ with a random variable $\tilde{\psi}_j$, taking values in $\mathcal{X}$. For every x, the $\tilde{\psi}_j$'s are reordered based on their distance to x, and this ordering is then used to define a permutation of $(v_j, \tilde{\theta}_j)$. They successfully apply this idea to stochastic volatility and spatial modeling but do not discuss how to handle discrete covariates.

Dunson and Park [2008] developed a kernel stick-breaking approach, which defines

$$v_j(x) = v_j K(x; \tilde{\psi}_j),$$

for some kernel on $\mathcal{X}$ with parameter $\tilde{\psi}_j$ such that $0 \leq K(x; \tilde{\psi}_j) \leq 1$. Dunson and Park use this approach for an application in epidemiology, and Reich and Fuentes [2007] apply the idea to a spatial dataset concerning hurricane wind fields. In the first application, the squared exponential kernel is used, so that

$$v_j(x) = v_j \exp(-\tilde{\tau}_j ||x - \tilde{\mu}_j||^2).$$

While, in the second, the authors consider both the squared exponential kernel and the uniform kernel, where $v_j(x)$ is defined as

$$v_j(x) = v_j \prod_{h=1}^{p} \mathbf{1}(|x_h - \tilde{\mu}_{j,h}| < \tilde{\tau}_j^{-1}).$$

Both examples involve continuous covariates only, and to incorporate discrete covariates, adequate kernels must be specified.

Two closely related approaches are given in Chung and Dunson [2011] and Griffin and Steele [2010]. In the first approach, the kernel is defined as the indicator that x lies in a ball of radius of r around $\tilde{\psi}_j$, i.e.

$$v_j(x) = v_j \mathbf{1}(||x - \tilde{\psi}_j|| < r).$$

The later extends this idea by defining the kernel as the indicator that x lies in a random subset $\tilde{\psi}_j$ of $\mathcal{X}$, i.e.

$$v_j(x) = v_j \mathbf{1}(x \in \tilde{\psi}_j).$$

Another common method defines the covariate-dependent stick length proportions by extending ideas in generalized linear models. In this case,

$$v_j(x) = l(\tilde{\psi}_j(x)),$$

where $l : \mathbb{R} \to [0, 1]$ is a monotone, differentiable link function and $\tilde{\psi}_j(x)$ is a random, real-valued function on $\mathcal{X}$. The function $l(\cdot)$ is commonly chosen to be the probit or logit link function, and $\tilde{\psi}_j(x)$ may be defined as a simple linear function, as a linear combination of basis functions, or through Gaussian process prior. For example, Rodriguez and Dunson [2011] use a probit link function and consider four possibilities for $\tilde{\psi}_j(x)$ depending on the application at hand: 1) for classic regression problems with continuous covariates, $\tilde{\psi}_j(\cdot)$ has a Gaussian process prior with a constant mean and the squared exponential covariance function; 2) for spatial and temporal applications, $\tilde{\psi}_j(\cdot)$ is a Gaussian Markov random field; 3) for discrete covariates, $\tilde{\psi}_j(\cdot)$ has a multivariate Gaussian distribution with a constant mean and identity covariance matrix; 4) in applications with both

continuous and discrete covariates, they assume $\tilde{\psi}_j(x)$ is a linear function of the continuous covariates with slopes that depend on the value of the discrete covariates. Chung and Dunson [2009] also use a probit link function but assume that $\tilde{\psi}_j(x)$ is a linear function of the absolute value of x. Ren et al. [2011] employ a logistic link function and basis function expansion of $\tilde{\psi}_j(x)$ in terms of squared exponential basis functions. Pati et al. [2012] study a probit link function and a zero mean Gaussian process prior for $\tilde{\psi}_j(x)$ with squared exponential covariance functions whose bandwidth depends on j. Applications in Rodriguez and Dunson [2011], Chung and Dunson [2009], and Ren et al. [2011] include stochastic volatility models, epidemiological studies, and image segmentation.

When y is continuous and univariate, the kernel for y is typically the standard linear regression kernel. Other response types require replacing the linear regression model with an appropriate kernel. For example, if the response is binary, ordinal, categorical, or counts, a generalized linear model seems appropriate.

Consistency

Posterior consistency results were recently studied for the kernel and probit stick-breaking models, where the notion of consistency is equivalent to the ideas used for models with covariate-dependent atoms in Section 2.3.3. Norets and Pelenis [2012b] study the former with the kernel defined by

$$K(x;\tilde{\psi}_j) = K(-\tilde{\tau}_j||x-\tilde{\mu}_j||^2),$$

where $K(\cdot)$ is continuous, is non-decreasing, has bounded derivative on $(-\infty, 0]$, and satisfies $0 < K(-z) < 1$ for $z \in [0,\infty)$, with additional reasonable conditions on the behavior of $K(-z)$ as $z \to \infty$. For example, $K(-z) = \exp(-z)$ satisfies their conditions. The response is assumed to be univariate and continuous, and the kernel for y is a scale-location density with additional constraints that are satisfied by the commonly chosen normal density. Weak consistency is demonstrated for a large class of conditional densities with minor conditions on the support of $\tilde{\theta}_j$. Strong consistency requires additional constraints on the prior of $\tilde{\theta}_j$, which are

satisfied by the normal and inverse-gamma priors that are used in practice, and on the priors of $\tilde{\psi}_j$ and v_j. In particular, they require a large prior mass on values of v_j close to 1. This is because given v_j and $\tilde{\psi}_j$, v_j is the maximum value of $w_j(x)$ for any x. Thus, in order for the weights to be able to peak close to one, the prior mass on values of v_j close to 1 must be large.

The latter, posterior consistency of the probit stick-breaking model, is studied by Pati et al. [2012], who prove weak and strong consistency for a large class of conditional densities with $v_j(x)$ defined by i.i.d. realizations of Gaussian processes through a probit link function. Again, the response is assumed to be univariate and continuous, and the kernel for y is the normal linear regression model. For weak consistency, continuity and approximation properties are required for the Gaussian processes. For strong consistency, the Gaussian processes must satisfy additional constraints. In particular, they assume

$$\begin{aligned}&\tilde{\psi}_j(x)|\tilde{\tau}_j \overset{iid}{\sim} \mathrm{GP}(0, c),\\ &c(x_1, x_2) = c\exp(-\tilde{\tau}_j||x_1 - x_2||^2),\end{aligned}$$

where the random bandwidths, $\tilde{\tau}_j$, are required to decay to zero at a fast enough rate, so that dependence on x in the weights decays with increasing j. From a computational perspective, for the probit stick-breaking approach, computations can be performed by introducing latent normal variables, but the number of latent variables that need to be updated can be huge. The kernel stick break approach has the advantage that $v_j(\cdot)$ is defined through a finite dimensional parameter $\tilde{\psi}_j$ and a known function, so that the numbers of computations is much more reasonable.

2.3.5 Other approaches

Another important class of models extends the random partition model and urn scheme of the DP to depend on covariates. For these models, obtaining a representation in terms of (2.5) can be far from straightforward. Reversely, deriving an expression for the random partition model and urn

scheme induced by (2.5) can also be difficult. An exception is when the random partition model and urn scheme correspond to the joint model (see Park and Dunson [2010]), then deriving a representation in terms of (2.5) is straightforward, and vice versa.

Müller and Quintana [2010] develop a general class of covariate-dependent random partition models defined by

$$p(\rho_n|x_{1:n}) \propto \prod_{j=1}^{k} c(S_j)g(x_j^*),$$

where $S_j = \{i \in \{1,\ldots,n\} : s_i = j\}$ and $x_j^* = \{x_i\}_{i\in S_j}$. The term $c(S_j)$ is called the cohesion function, and for example, $c(S_j) = \Gamma(n_j)$ for the DP. The similarity function, $g(\cdot)$, captures the closeness of covariates, where large values indicate high similarity. The covariate-dependent random partition model of the joint approach is a special case, satisfies marginalization and scalability properties, and is easier from a computational perspective; thus, in examples, it is their focus. In Müller et al. [2012], the covariate-dependent random partition model is extended to allow variable selection.

Proposals that modify the urn scheme to depend on the covariates include Rasmussen and Ghahramani [2002], Dahl [2008], and Blei and Frazier [2011], just to mention a few. In most cases, the probability that a new subject is allocated to j^{th} *cluster* is altered to depend on the covariates in that cluster, so that

$$p(s_{n+1}|s_{1:n}, x_{1:n+1}) \propto \begin{cases} g(x_{n+1}|x_j^*) & \text{if } s_{n+1} = j \\ \alpha & \text{if } s_{n+1} = k+1 \end{cases}.$$

The function $g(x_{n+1}|x_j^*)$ is a measure of the similarity of x_{n+1} and the covariates in the j^{th} cluster and may be defined through a distance (Dahl [2008], Blei and Frazier [2011]) or kernel function (Rasmussen and Ghahramani [2002]).

In Dunson et al. [2007], the random covariate-dependent probability measure $\mathbf{P}_x$ is defined through a weighted mixture of n independent random probability measures with weights constructed through kernel func-

tions centered at the observed covariate values:

$$\mathbf{P}_x = \sum_{i=1}^{n} \frac{w_i K(x; x_i)}{\sum_{i'=1}^{n} w_{i'} K(x; x_{i'})} \mathbf{P}_i,$$

where $P_i \overset{iid}{\sim} \text{DP}(\alpha P_0)$. However, because the prior of $\mathbf{P}_x$ depends on the sample size and observed covariates, it is unappealing from a Bayesian perspective and lacks desirable marginalization and updating properties (see Dunson [2010] for more details).

Other proposals along the lines of (2.5) focus exclusively on discrete categorical covariates, where, for example, x might indicate the hospital, among M hospitals, where the patient was treated. An interesting proposal for the law of $\mathbf{P}_x$, in this setting, is the hierarchical Dirichlet process of Teh et al. [2006], who assume $\mathbf{P}_x \mid P_0 \overset{iid}{\sim} \text{DP}(\alpha P_0)$ and model the random base measure $\mathbf{P}_0$ nonparametrically, where $\mathbf{P}_0 \sim \text{DP}(\gamma H)$. A further development is the nested Dirichlet process (Rodriguez and Dunson [2011]) where the model is given as $\mathbf{P}_x \mid Q \overset{iid}{\sim} Q$ and $\mathbf{Q} \sim \text{DP}(\alpha \text{DP}(\gamma H))$. Alternative proposals are given by Müller et al. [2004], Walker and Muliere [2003], Kolossiatis et al. [2011], Griffin et al. [2011], and Lijoi et al. [2011], just to mention a few. In this setting, x is just a label and the distance between two covariate values has no meaning. This will not be our focus.

2.4 Summary

In summary, there are three main types of models used in practice for covariate-dependent density estimation through nonparametric mixture models: 1) models based on the joint approach (2.4); 2) models based on the conditional approach with constant weights and flexible atoms (2.8); and 3) models based on the conditional approach with flexible weights and simple atoms (2.11). Another important class is comprised of models based on covariate-dependent random partition models or urn schemes. In a specific case, such models correspond to the joint model (2.4), but in general, they are in the flavor of models with flexible weights and simple atoms (2.11).

Across model type, there are advantages and disadvantages. The joint model is flexible and has the advantage of computational simplicity, but, modelling of x is required, even though interest is only in the conditional of y given x. The drawbacks of this will be discussed in detail in Chapter 4. The conditional approach, on the other hand, has the advantage of modelling the conditional directly, which can lead to improved estimates. When constant weights are assumed, computations can be relatively easy. However, in order to capture a wide range of data-generating conditional densities, the atoms must be very flexible, which can greatly increase the computational burden. Furthermore, with increasing flexibility in the atoms, interpretations become increasingly difficult. The conditional approach with covariate-dependent weights tends to be very flexible but can be computational burdensome. Interpretations can also be hard.

Within each model type, the number of model and prior choices is large, and deciding among them can be challenging.

For the practical purposes of defining a model for a given dataset, a detailed study of model properties is needed both within and across model types. Consistency studies provide an interesting validation of the models, but the types of models under study are extremely flexible, and it is likely that most are consistent. In the remaining chapters, our aim is to carefully examine properties of the various models and priors of interest and the effects of these properties on prediction.

Chapter 3

Enriched Dirichlet process

In Bayesian nonparametric mixture models, the Dirichlet process is quite often used as a prior for the mixing measure, and, typically, the mixing parameter is multivariate, so that the Dirichlet process is a prior on the set of probability measures on $\mathbb{R}^p$, $p > 1$. In this setting, however, a Dirichlet process prior can be restrictive in the sense that the variability is determined by a single parameter α, regardless of p. The aim of this chapter is to highlight this drawback and to construct an enrichment of the Dirichlet process that is more flexible with respect to the precision parameter yet still conjugate, starting from the notion of enriched conjugate priors, which address an analogous lack of flexibility of standard conjugate priors in a parametric setting. Properties of the resulting enriched conjugate nonparametric prior are discussed in detail including an urn scheme and stick-breaking representation. Finally, we consider an application to mixture models that allows for uncertainty between homoskedasticity and heteroskedasticity. In Chapter 4, this process will be utilized to define a novel Bayesian nonparametric regression model.

This chapter is joint work with Silvia Mongelluzzo and Sonia Petrone

and is based on Wade et al. [2011], which was awarded the 2010 Lindley prize by the International Society of Bayesian Analysis.

3.1 Motivation

Conjugacy is a desirable property because the posterior distribution remains analytically tractable; this is especially true in nonparametric inference where the posterior distribution of non-conjugate priors can be very complex. The most popular prior in Bayesian nonparametric inference is the Dirichlet process, and it is conjugate; if $Z_i \mid \mathbf{P} = P$ are independent and identically distributed (i.i.d.) according to P, and $\mathbf{P}$ is a Dirichlet process, $\mathrm{DP}(\alpha P_0)$, with precision parameter α and base measure P_0 on the sample space $\mathcal{Z}$, then

$$\mathbf{P} \mid Z_1 = z_1, \ldots, Z_n = z_n \sim \mathrm{DP}(\alpha P_0 + \sum_{i=1}^{n} \delta_{z_i}).$$

However, when Z is a random vector and $\mathbf{P}$ is a random probability measure on $\mathbb{R}^p$, $p > 1$, as in many applications including regression settings, the choice of a Dirichlet process prior implies that the variability is determined by a single parameter α. Indeed, the precision parameter α plays an important role; it not only reflects the strength of belief in the prior guess of P_0, but also controls the ties configuration in a random sample from $\mathbf{P}$. Thus, having only one degree of freedom, α, in the prior can be quite restrictive.

In fact, a similar lack of flexibility arises in a parametric setting; standard conjugate priors for the natural exponential family have only one parameter to control variability. To overcome this issue, a general class of *enriched conjugate priors* (Consonni and Veronese [2001]) have been proposed. A Dirichlet process, $\mathrm{DP}(\alpha P_0)$, is characterized by the fact that the finite dimensional distributions of the probability over any measurable partition,$(C_1, \ldots, C_m)$, of $\mathcal{Z}$, are Dirichlet with parameters $(\alpha P_0(C_1), \ldots, \alpha P_0(C_m))$. The Dirichlet process inherits conjugacy from the property of conjugacy of the standard Dirichlet distribution prior for

multinomial sampling, but also inflexibility from the fact that the Dirichlet distribution, as all standard conjugate priors, has only one parameter to control variability. The question addressed in this chapter is whether one can extend the notion of enriched conjugate priors to nonparametric inference and construct a prior on a random probability measure over $\mathbb{R}^p$, that is more flexible than the DP in allowing more parameters to control the variability, yet is still conjugate.

Actually, Doksum's Neutral to the Right Process (Doksum [1974]) is an extension of the enriched conjugate Generalized Dirichlet distribution to a process, providing a more flexible, conjugate prior for *univariate* random distribution functions. The Generalized Dirichlet distribution is defined for a specific ordering of the random probabilities; thus, extension to a multivariate random distribution is not obvious, since there is no natural ordering in $\mathbb{R}^p$.

Therefore, we start our analysis by constructing an enriched Dirichlet prior for a multivariate random distribution when the sample space is finite. To convey the main ideas, we will focus on the case when the random vector Z can be partitioned into two groups, $Z = (X, Y)$, and the sample space can be written as the product of two finite spaces (or in the more general case, the product of two complete separable metric spaces, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$). In the finite case, the enriched Dirichlet distribution is obtained based on the reparametrization of the joint probabilities in terms of the marginal and the conditionals.

Then, we extend this construction to a process by reparametrizing the joint random probability measure in terms of the marginal and conditionals and assigning independent Dirichlet process priors to each of these terms. The parameters of the resulting *enriched* Dirichlet process again include a base measure controlling the location, but there are now many more parameters to control the variability. We show that the Dirichlet process is in fact a special case, which consequently, characterizes the distribution of the random conditionals. Although many desirably properties are maintained, some are necessarily weakened, including a clear asymmetry in the two (groups of) variables, that however may be reasonable in

several applications.

Applications to mixture models involve simply replacing the sample space $\mathcal{Z}$ with the parameter space. Extensions for Bayesian nonparametric regression through mixture models are developed in Chapter 4.

The remainder of this chapter is organized as follows. In Section 3.2, we give a brief overview of enriched conjugate priors for the natural exponential family. In Section 3.3, we discuss the *enriched* Dirichlet distribution in the finite case as a particular enriched conjugate prior for multinomial sampling and provide a Pólya urn characterization. These notions are extended to a process in Section 3.4. Finally, a simple application to mixture models is illustrated using data on national test scores to compare schools in Section 3.5.

3.2 Preliminaries: enriched conjugate priors

For a Natural Exponential Family (NEF) $\mathcal{F}$ on $\mathbb{R}^d$, where d represents the dimension of the sufficient statistics, the likelihood for the natural parameter θ is given by

$$L_\theta(\theta|s,n) = \exp(\theta' s - nM(\theta)) \quad \text{for } \theta \in \Theta,$$

where s is a d-dimensional vector of the sufficient statistics,

$$M(\theta) = \log \int \exp(\theta' x)\eta(dx),$$

and η is a σ-finite measure on the Borel sets of $\mathbb{R}^d$. The parameter space Θ is the interior of the set $\mathcal{N} = \{\theta \in \mathbb{R}^d : M(\theta) < \infty\}$. More generally, we have a Standard Exponential Family (SEF) if $\Theta \subseteq \mathcal{N}$, and it is non-empty and open.

A family of measures on the Borel sets of Θ whose densities with respect to the Lebesgue measure are of the form

$$\pi_\theta(\theta|s^*,n^*) \propto L_\theta(\theta|s^*,n^*)$$

is called the *standard conjugate family of priors* of $\mathcal{F}$ relative to the parametrization θ, where the sufficient statistics, s, are replaced by parameters, s^*, which control the location of the prior, and the sample size, n, is replaced by a single parameter, n^*, which controls the precision; see Diaconis and Ylvisaker [1979].

Consonni and Veronese [2001] discuss *enriched* conjugate priors for the NEF, moving from the notion of conditional reducibility. A d-dimensional NEF is called k *conditionally reducible* if the density can be decomposed as the product of k standard exponential families, each depending on their own parameters. The notion of *enriched conjugate priors* involves replacing the sufficient statistics and the sample size with different hyperparameters within each SEF. This means giving independent standard conjugate priors to the parameters of the conditional densities and induces a prior on the original parameter of the NEF which enriches the standard conjugate prior by allowing for k precision parameters. For a deeper discussion, see Consonni and Veronese [2001].

One important example is given by the Generalized Dirichlet distribution of Connor and Mosimann [1969], which provides an enriched conjugate prior for the parameters of a multinomial distribution; see Consonni and Veronese [2001], Example 4. Briefly, if $(N_1, \ldots, N_k)$ is multinomial given $(\mathbf{p}_1 = p_1, \ldots, \mathbf{p}_k = p_k)$, one can decompose the multinomial probability function as

$$\begin{aligned} p(n_1, \ldots, n_k \mid p_1, \ldots, p_k) =& p(n_1 \mid v_1) p(n_2 \mid n_1, v_2) \\ & * \ldots * p(n_k \mid n_1, \ldots, n_{k-1}, v_k), \end{aligned}$$

where each factor in the product is a NEF (namely, binomial), depending on its own parameter,

$$\begin{aligned} \mathbf{v}_1 &= \mathbf{p}_1, \\ \mathbf{v}_i &= \mathbf{p}_i / (1 - \sum_{j=1}^{i-1} \mathbf{p}_j) \quad \text{for } i = 2, \ldots, k-1, \end{aligned}$$

and $\mathbf{v}_k$ is degenerate at 1, which guarantees $\sum_{j=1}^{k} \mathbf{p}_j = 1$ a.s. The stan-

dard, Dirichlet$(\alpha_1, \ldots, \alpha_k)$ conjugate prior corresponds to assuming

$$\mathbf{v}_i \overset{ind}{\sim} \text{Beta}(\alpha_i, \sum_{j=i+1}^{k} \alpha_j) \quad \text{for } i = 1, \ldots, k-1.$$

The enriched, or Generalized, Dirichlet conjugate prior allows a more flexible choice of the beta hyperparameters;

$$\mathbf{v}_i \overset{ind}{\sim} \text{Beta}(\alpha_i, \beta_i) \quad \text{for } i = 1, \ldots, k-1.$$

It is worth underlining that some properties of the Dirichlet distribution are necessarily weakened. In particular, the Dirichlet prior implies that *any permutation* of $(\mathbf{p}_1, \ldots, \mathbf{p}_k)$ is completely neutral (the vector $(\mathbf{p}_1, \ldots, \mathbf{p}_k)$ is completely neutral if and only if $(\mathbf{p}_1, \mathbf{p}_2/(1-\mathbf{p}_1), ..., \mathbf{p}_k/(1-\sum_{j=1}^{k-1} \mathbf{p}_j))$ are independent). The Generalized Dirichlet only assumes that *one* ordered vector $(\mathbf{p}_1, \ldots, \mathbf{p}_k)$ is completely neutral. This makes applications to the bivariate case of contingency tables $\mathbf{p}_{i,j}$ not obvious, since there is no natural ordering in two dimensions. The enriched conjugate prior that we propose in the next section is a simple proposal in this direction.

3.3 Finite case: enriched Dirichlet distribution

Let $\{(X_n, Y_n)\}_{n \in \mathbb{N}}$ be a sequence of discrete random vectors with values in $\mathcal{X} \times \mathcal{Y} = \{1, \ldots, k\} \times \{1, \ldots, m\}$, such that $(X_i, Y_i) \mid \mathbf{p} = p \overset{iid}{\sim} p$, where $\mathbf{p}$ is a random probability function with mass $\mathbf{p}_{i,j}$ on (i, j), $i = 1, \ldots, k$; $j = 1, \ldots, m$. Then, given $\mathbf{p} = p$, the vector of counts $(N_{1,1}, \ldots, N_{k,m})$, where $N_{i,j}$ is the number of times the pair (i, j) is observed in a sam-

ple $((X_1, Y_1), \ldots, (X_n, Y_n))$, has a multinomial probability function;

$$p(n_{1,1}, ..., n_{k,m-1} \mid p_{1,1}, ..., p_{k,m-1}) = \frac{n!}{n_{1,1}!...n_{k,m-1}!(n - \sum_{(i,j)\neq(k,m)} n_{i,j})!}$$
$$* \, p_{1,1}^{n_{1,1}} \cdots p_{k,m-1}^{n_{k,m-1}} \, (1 - \sum_{(i,j)\neq(k,m)} p_{i,j})^{n - \sum_{(i,j)\neq(k,m)} n_{i,j}}, \tag{3.1}$$

for $n_{i,j} \geq 0$; $\sum_{i=1}^{k} \sum_{j=1}^{m} n_{i,j} = n$. The standard conjugate prior for $(\mathbf{p}_{1,1}, \ldots, \mathbf{p}_{k,m})$ is the Dirichlet distribution, which involves replacing the $km-1$ sufficient statistics in (3.1) with hyperparameters, $s^* = (s^*_{1,1}, ..., s^*_{k,m-1})$, that control the location of the prior, and the sample size with a single hyperparameter, n^*, that controls the precision of the prior. As discussed in Section 3.2, a generalized Dirichlet prior is problematic in this case, since there is no natural ordering of the probabilities $\mathbf{p}_{i,j}$.

However, a fairly natural and simple enrichment can be obtained by first applying the linear transformation

$$N_{i+} = \sum_{j=1}^{m} N_{i,j} \quad \text{for } i = 1, ..., k-1,$$
$$N_{i,j} = N_{i,j} \quad \text{for } i = 1, ..., k \quad j = 1, ..., m-1,$$

followed by the reparametrization

$$\mathbf{p}_{i+} = \sum_{j=1}^{m} \mathbf{p}_{i,j} \quad \text{for } i = 1, ..., k-1,$$
$$\mathbf{p}_{j|i} = \frac{\mathbf{p}_{i,j}}{\mathbf{p}_{i+}} \quad \text{for } i = 1, ..., k-1 \quad j = 1, ..., m-1,$$
$$\mathbf{p}_{j|k} = \frac{\mathbf{p}_{k,j}}{1 - \sum_{i=1}^{k-1} \mathbf{p}_{i+}} \quad \text{for } j = 1, ..., m-1.$$

Define: $\underline{N}_+ = (N_{1+}, ..., N_{k-1+})$; $\underline{N}_i = (N_{i,1}, ..., N_{i,m-1})$; $\underline{\mathbf{p}}_+ = (\mathbf{p}_{1+}, ..., \mathbf{p}_{k-1+})$, and $\underline{\mathbf{p}}_i = (\mathbf{p}_{1|i}, ..., \mathbf{p}_{m-1|i})$, for $i = 1, ..., k$. Under this linear transformation and reparametrization, the multinomial is a $k+1$ conditionally

reducible NEF;

$$p(\underline{n}_+, \underline{n}_1, ..., \underline{n}_k \mid \underline{p}_+, \underline{p}_1, ..., \underline{p}_k) = p(\underline{n}_+ \mid \underline{p}_+) \prod_{i=1}^{k} p(\underline{n}_i \mid \underline{p}_i, \underline{n}_+), \tag{3.2}$$

$$(N_{i,1}, ..., N_{i,m} \mid n_{i+}, p_{1|i}, ..., p_{m|i}) \sim \text{Mult}\left(n_{i+}, p_{1|i}, ..., p_{m|i}\right) \quad \text{for } i = 1, ..., k,$$
$$(N_{1+}, ..., N_{k+} \mid p_{1+}, ..., p_{k+}) \sim \text{Mult}(n, p_{1+}, ..., p_{k+}).$$

By replacing the sufficient statistics and sample size with different parameters within each SEF in the right hand side of (3.2), one can create a more flexible conjugate prior. In particular, letting $(s^*_{(+)}, s^*_{(1)}, ..., s^*_{(k)})$ denote the $km - 1$ location parameters and $(n^*_+, n^*_1, ..., n^*_k)$ denote the precision parameters, in terms of $(\underline{\mathbf{p}}_+, \underline{\mathbf{p}}_1, ..., \underline{\mathbf{p}}_k)$, the Enriched Dirichlet conjugate prior is

$$\mathbf{p}_{1+}, ..., \mathbf{p}_{k+} \sim \text{Dir}(s^*_{1+}, ..., s^*_{k-1+}, n^*_+ - \sum_{i=1}^{k-1} s^*_{i+}), \tag{3.3}$$
$$\mathbf{p}_{1|i}, ..., \mathbf{p}_{m|i} \sim \text{Dir}(s^*_{i,1}, ..., s^*_{i,m-1}, n^*_i - \sum_{j=1}^{m-1} s^*_{i,j}),$$

where $(\mathbf{p}_{1+}, ..., \mathbf{p}_{k+}), (\mathbf{p}_{1|1}, ..., \mathbf{p}_{m|1}), ..., (\mathbf{p}_{1|k}, ..., \mathbf{p}_{m|k})$ are independent. We get back to the Dirichlet distribution if $n^*_i = s^*_{i+}$ for $i = 1, ...k - 1$ and $n^*_+ = \sum_{i=1}^{k} n^*_i$.

Remark 1. The Dirichlet distribution on the vector $\underline{\mathbf{p}} = (\mathbf{p}_{1,1}, ..., \mathbf{p}_{k,m})$ defining the random marginal, $\mathbf{p}_x$, $\mathbf{p}_y$, and conditional, $\mathbf{p}_{y|x}$, $\mathbf{p}_{x|y}$, probability functions is characterized by the properties

(i) $\mathbf{p}_x(\cdot)$ and $\mathbf{p}_{y|x}(\cdot|i), i = 1, \ldots, k$ are independent, and

(ii) $\mathbf{p}_y(\cdot)$ and $\mathbf{p}_{x|y}(\cdot|j), j = 1, \ldots, m$ are independent;

see Geiger and Heckerman [1997]. The Enriched Dirichlet relaxes that the independence properties holds in both directions. We maintain (i) and allow more degrees of freedom in the distributions of $\mathbf{p}_x$ and $\mathbf{p}_{y|x}$.

Remark 2. Under the linear transformation discussed here, the multinomial could also be viewed as a $km-1$ conditionally reducible NEF; it can be written as the product of $km-1$ SEFs (namely, binomial) each depending on its own parameters. The resulting enriched conjugate prior has $km-1$ parameters to control the precision and can be seen as nested version of Generalized Dirichlet distribution of Connor and Mosimann [1969].

In the rest of the chapter, we will use the following parametrization of the distributions (3.3). Let $\alpha(\cdot)$ be a finite measure on $\mathcal{X}$ and $\mu(\cdot,\cdot)$ be a mapping from $2^{\mathcal{Y}} \times \mathcal{X}$ to $\mathbb{R}_+$ such that for every $x \in \mathcal{X}$, $\mu(\cdot, x)$ is a finite measure on $(\mathcal{Y}, 2^{\mathcal{Y}})$. Then we assume that the parameters in (3.3) are chosen in terms of $\alpha(\cdot)$ and $\mu(\cdot,\cdot)$;

$$\begin{aligned} &\mathbf{p}_{1+}, ..., \mathbf{p}_{k+} \sim \mathrm{Dir}(\alpha(1), ..., \alpha(k)), \\ &\mathbf{p}_{1|i}, ..., \mathbf{p}_{m|i} \sim \mathrm{Dir}(\mu(1,i), ..., \mu(m,i)) \quad i = 1, ..., k, \end{aligned} \tag{3.4}$$

with the convention that if $\alpha(i) = 0$ then $\mathbf{p}_{i+}$ is degenerate at 0 and if $\mu(j,i) = 0$ then $\mathbf{p}_{j|i}$ is degenerate at 0. If $\alpha(i) > 0$ and $\mu(j,i) > 0$ for all i, j, then the enriched Dirichlet conjugate prior induced on $(\mathbf{p}_{1,1}, \ldots, \mathbf{p}_{k,m})$ is

$$\begin{aligned} &f(p_{1,1}, ..., p_{k,m-1}) \\ &= \frac{\Gamma(\alpha(\mathcal{X}))}{\prod_{i=1}^{k} \Gamma(\alpha(i))} \prod_{i=1}^{k-1} (\sum_{j=1}^{m} p_{i,j})^{\alpha(i)-\mu(\mathcal{Y},i)} (1 - \sum_{i=1}^{k-1} \sum_{j=1}^{m} p_{i,j})^{\alpha(k)-\mu(\mathcal{Y},k)} \\ &* \prod_{i=1}^{k} \frac{\Gamma(\mu(\mathcal{Y},i))}{\prod_{j=1}^{m} \Gamma(\mu(j,i))} \prod_{j=1}^{m-1} p_{i,j}^{\mu(j,i)-1} \prod_{i=1}^{k-1} p_{i,m}^{\mu(m,i)-1} (1 - \sum_{(i,j)\neq(k,m)} p_{i,j})^{\mu(m,k)-1}. \end{aligned}$$

Clearly, the prior of the marginal probabilities $(\mathbf{p}_{+1}, \ldots, \mathbf{p}_{+m})$ on $\mathcal{Y}$ is no longer a Dirichlet distribution, and in fact, the density may not be available in closed form. But, we can give the following representation in terms of G-Meijer variables (Springer and Thompson [1970]). First, remembering the Gamma representation of the Dirichlet distribution and defining $\mathbf{v}_i \overset{ind}{\sim} \mathrm{Gamma}(\alpha(i), 1)$ and $\mathbf{v}_{ij} \overset{ind}{\sim} \mathrm{Gamma}(\mu(j,i), 1)$, we have the

following G-Meijer representation of the vector $(\mathbf{p}_{1,1}, ..., \mathbf{p}_{k,m})$

$$(\mathbf{p}_{1,1}, ..., \mathbf{p}_{k,m}) \overset{d}{=} \left(\frac{\mathbf{v}_1 \mathbf{v}_{11}}{\sum_{i=1}^{k} \mathbf{v}_i \sum_{j=1}^{m} \mathbf{v}_{1j}}, ..., \frac{\mathbf{v}_k \mathbf{v}_{km}}{\sum_{i=1}^{k} \mathbf{v}_i \sum_{j=1}^{m} \mathbf{v}_{kj}} \right),$$

which is independent of $\sum_{i=1}^{k} \mathbf{v}_i \sum_{j=1}^{m} \mathbf{v}_{1j}, \ldots, \sum_{i=1}^{k} \mathbf{v}_i \sum_{j=1}^{m} \mathbf{v}_{kj}$; where the symbol $\overset{d}{=}$ denotes equality in distribution. Therefore, the marginal probabilities over $\mathcal{Y}$ can be represented as the sum of G-Meijer random variables;

$$(\mathbf{p}_{+1}, ..., \mathbf{p}_{+m}) \overset{d}{=} \left(\sum_{i=1}^{k} \frac{\mathbf{v}_i \mathbf{v}_{i1}}{\sum_{h=1}^{k} \mathbf{v}_h \sum_{j=1}^{m} \mathbf{v}_{ij}}, ..., \sum_{i=1}^{k} \frac{\mathbf{v}_i \mathbf{v}_{im}}{\sum_{h=1}^{k} \mathbf{v}_h \sum_{j=1}^{m} \mathbf{v}_{ij}} \right).$$

3.3.1 Enriched Pólya urn

An alternative way to define the Enriched Dirichlet distribution is based on a Pólya urn scheme, which will be useful in extending the distribution to a process. In the bivariate setting, the standard Pólya urn scheme describes the predictive distribution of a sequence of random vectors. An urn contains pairs of balls of color $(i, j) \in \mathcal{X} \times \mathcal{Y}$. A pair of balls is drawn from the urn and replaced along with another pair of balls of the same colors. The random vector, (X_n, Y_n), is equal to (i, j) if the n-th pair drawn is of color (i, j).

Alternatively, we can consider one urn containing just X-balls and k urns, say $Y|i$ urns, containing only Y-balls. We first draw an X-ball from the X-urn and replace it along with another ball of the same color, and then, depending on color of the X-ball, draw a Y-ball from urn associated to X-ball drawn, and replace it along with another ball of the same color. In this case, the random vector, (X_n, Y_n), is equal to (i, j) if the n-th X-ball drawn is of color i and the Y ball associated with it is of color j. If the number of Y-balls in the $Y|i$ urn is equal to the number balls of color i in the X-urn, the two urn schemes are equivalent.

The Enriched Pólya Urn scheme enriches this urn scheme by relaxing the constraints that the number of Y-balls in the $Y|i$ urn has to equal the

number of X-balls of color i in the X-urn for $i = 1, ..., k$. More precisely, the number of balls in each urn is specified as follows:

- $\alpha(i)$ is the number of X-balls of color i
- $\mu(j, i)$ is the number of Y-balls of color j in the $Y|i$ urn

where $\alpha(\mathcal{X}) = \sum_{i=1}^{k} \alpha(i)$ is the total number of balls in the X-urn and $\mu(\mathcal{Y}, i) = \sum_{j=1}^{m} \mu(j, i)$ is the total number of balls in the $Y|i$ urn for $i = 1, ..., k$. This urn scheme implies the following predictive distribution:

$$Pr(X_1 = i, Y_1 = j) = \frac{\alpha(i)}{\alpha(\mathcal{X})} \frac{\mu(j, i)}{\mu(\mathcal{Y}, i)},$$

$$Pr(X_{n+1} = i, Y_{n+1} = j | X_1 = i_1, Y_1 = j_1, .., X_n = i_n, Y_n = j_n)$$
$$= \frac{\alpha(i) + \sum_{h=1}^{n} \delta_{i_h}(i)}{\alpha(\mathcal{X}) + n} \frac{\mu(j, i) + \sum_{h=1}^{n} \delta_{j_h, i_h}(j, i)}{\mu(\mathcal{Y}, i) + \sum_{h=1}^{n} \delta_{i_h}(i)}.$$

Theorem 3.3.1 *Let $\{(X_n, Y_n)\}_{n \in \mathbb{N}}$ be a sequence of random vectors taking values in $\{1, ..., k\} \times \{1, ..., m\}$ with predictive distributions characterized by an Enriched Pólya urn scheme with parameters $\alpha(\cdot)$ and $\mu(\cdot, \cdot)$. Then,*

1. *the sequence of random vectors $\{(X_n, Y_n)\}_{n \in \mathbb{N}}$ is exchangeable, and its de Finetti measure is an Enriched Dirichlet distribution with parameters $\alpha(\cdot)$ and $\mu(\cdot, \cdot)$.*
2. *as $n \to \infty$, the sequence of the predictive distributions $p_n(i, j) = Pr(X_{n+1} = i, Y_{n+1} = j | X_1 = i_1, Y_1 = j_1, .., X_n = i_n, Y_n = j_n)$ converges a.s with respect to the exchangeable law to a random probability function, $\mathbf{p}$; and $\mathbf{p}$ is distributed according to the Enriched Dirichlet de Finetti measure.*

Proof. The proof is an extension of that used for the standard Pólya urn (see Ghosh and Ramamoorthi [2003], pages 94-95). The first step is to show the sequence of random vectors is exchangeable. Next, computing their finite dimensional distributions and using de Finetti's Representation Theorem, the random vectors are shown to be i.i.d given the

random variables $(\mathbf{p}_{1+}, ...\mathbf{p}_{k+}, \mathbf{p}_{1|1}, ..., \mathbf{p}_{m|k}) = (p_{1+}, ...p_{k+}, p_{1|1}, ..., p_{m|k})$ which are distributed according to an Enriched Dirichlet distribution with parameters α and μ.

From the predictive distribution, it follows that the joint distribution can be expressed as:

$$Pr(X_1 = i_1, Y_1 = j_1, ..., X_n = i_n, Y_n = j_n) = \prod_{l=1}^{n} \frac{\alpha(i_l) + \sum_{h=1}^{l-1} \delta_{i_h}(i_l)}{\alpha(\mathcal{X}) + l - 1} * \frac{\mu(j_l, i_l) + \sum_{h=1}^{l-1} \delta_{j_h, i_h}(j_l, i_l)}{\mu(\mathcal{Y}, i_l) + \sum_{h=1}^{l-1} \delta_{i_h}(i_l)},$$

which can be equivalently expressed as:

$$\frac{\Gamma(\alpha(\mathcal{X}))}{\prod_{i=1}^{k} \Gamma(\alpha(i))} \frac{\prod_{i=1}^{k} \Gamma(\alpha(i) + n_{i+})}{\Gamma(\alpha(\mathcal{X}) + n)} * \prod_{i=1}^{k} \frac{\Gamma(\mu(\mathcal{Y}, i))}{\prod_{j=1}^{m} \Gamma(\mu(j, i))} \prod_{i=1}^{k} \frac{\prod_{j=1}^{m} \Gamma(\mu(j, i) + n_{ij})}{\Gamma(\mu(\mathcal{Y}, i) + n_{i+})}. \tag{3.5}$$

The joint distribution only depends on the number of unique pairs seen, not on the order in which they are observed. Thus, the pairs $\{X_n, Y_n\}_{n \in \mathbb{N}}$ form an exchangeable sequence. By de Finetti's Representation Theorem, there exists a probability measure $\tilde{Q}$ on the simplex

$$S_{k,m} = \{p_{1,1}, ..., p_{k,m} : p_{i,j} \geq 0 \text{ and } \sum_{i=1}^{k} \sum_{j=1}^{m} p_{i,j} = 1\},$$

such that:

$$Pr(X_1 = i_1, Y_1 = j_1, ..., X_n = i_n, Y_n = j_n) = \int_{[0,1]^{km}} \prod_{i=1}^{k} \prod_{j=1}^{m} p_{i,j}^{n_{i,j}} \tilde{Q}(dp_{1,1}, ..., dp_{k,m}).$$

Define the simplexes

$$S_k = \{p_{1+}, ..., p_{k+} : p_{i+} \geq 0 \text{ and } \sum_{i=1}^{k} p_{i+} = 1\},$$

and

$$S_m^{(i)} = \{p_{i|1}, ..., p_{i|k} : p_{j|i} \geq 0 \text{ and } \sum_{j=1}^{m} p_{j|i} = 1\},$$

for $i = 1, ...k$. Let Q be the probability measure on the product of the simplexes $S_k \times \prod_{i=1}^{k} S_m^{(i)}$ obtained from $\tilde{Q}$ via a reparametrization in terms of $(\mathbf{p}_{1+}, ..., \mathbf{p}_{k+}, \mathbf{p}_{1|1}, ..., \mathbf{p}_{m|k})$. Then,

$$\begin{aligned} &Pr(X_1 = i_1, Y_1 = j_1, ..., X_n = i_n, Y_n = j_n) \\ &\quad = \int_{[0,1]^k \times [0,1]^{km}} \prod_{i=1}^{k} p_{i+}^{n_{i+}} \prod_{j=1}^{m} p_{j|i}^{n_{ij}} Q(dp_{1+}, ..., dp_{m|k}). \end{aligned} \tag{3.6}$$

Since the Dirichlet distribution is determined by its moments, combining equations (3.5) and (3.6) implies that

$$\begin{aligned} &\mathbf{p}_{1+}, ..., \mathbf{p}_{k+} \sim \text{Dir}(\alpha(1), ..., \alpha(k)), \\ &\mathbf{p}_{1|i}, ..., \mathbf{p}_{m|i} \sim \text{Dir}(\mu(1,i), ..., \mu(m,i)) \quad i = 1, ..., k, \end{aligned}$$

where $(\mathbf{p}_{1+}, ..., \mathbf{p}_{k+})$, $(\mathbf{p}_{1|1}, ..., \mathbf{p}_{m|1})$,..., and $(\mathbf{p}_{1|k}, ..., \mathbf{p}_{m|k})$ are independent.

The second part of the theorem follows from de Finetti's results on the asymptotic behavior of the predictive distributions for exchangeable sequences; see Cifarelli and Regazzini [1996]. ∎

3.4 Enriched Dirichlet process

Assume $\mathcal{X}$ and $\mathcal{Y}$ are complete and separable metric spaces with Borel σ-algebras $\mathcal{B}_X$ and $\mathcal{B}_Y$. Let $\mathcal{B}$ be the σ-algebra generated by the product of the σ-algebras of $\mathcal{X}$ and $\mathcal{Y}$ and $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ be the set of probability measures on the measurable product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B})$ where $\mathcal{M}(\mathcal{X})$, $\mathcal{M}(\mathcal{Y})$ are similarly defined. For any $P \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$, let P_X denote the marginal probability measure, $P_{Y|X}(\cdot|x)$ for $x \in \mathcal{X}$ denote a version of the conditional, and $P_{Y|X}$ denote the entire version of the conditional as an element of $\mathcal{M}(\mathcal{Y})^{\mathcal{X}}$. Here, we consider the Borel σ-algebra under weak convergence on $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$, $\mathcal{M}(\mathcal{X})$, and $\mathcal{M}(\mathcal{Y})$ and the product σ-algebra on $\mathcal{M}(\mathcal{Y})^{\mathcal{X}}$.

We will define a probability measure on $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ that is more flexible than the Dirichlet process with respect to the precision parameter and still retains conjugacy by extending the ideas of the Enriched Dirichlet distribution.

Note that trying to enrich the DP by using the Enriched Dirichlet in place of the Dirichlet as the finite dimensional distributions, i.e., defining a random $\mathbf{P}$ such that $(\mathbf{P}(A_1 \times B_1), \ldots, \mathbf{P}(A_k \times B_m)) \sim$ Enriched Dirichlet distribution, would not succeed because finite additivity holds only with a specification of the parameters that is equivalent to the Dirichlet distribution.

Instead, we use directly the idea of the Enriched Dirichlet distribution, which defines a prior for the joint by first, decomposing it in terms of the marginal and conditionals and then, assigning independent conjugate priors to them. If $\mathcal{X}, \mathcal{Y}$ are general spaces, it is a delicate issue to establish that such an approach induces a prior on the joint. In particular, given a prior on $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$, the map $(P_X, P_{Y|X}) \to \int_{(\cdot)} P_{Y|X}(\cdot|x) dP_X(x)$ induces a prior on $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ if it is jointly measurable in $(P_X, P_{Y|X})$, which is not true in general. Fortunately, if the prior for the marginal concentrates on the set of discrete probability measures and independence assumptions hold, the prior on the marginal and conditionals can be restricted to a subspace of $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$ that has measure one, and on this subspace, the mapping is measurable, which is shown after the following definition.

Definition 3.4.1 *Let α be a finite measure on $(\mathcal{X}, \mathcal{B}_X)$ and μ be a mapping from $(\mathcal{B}_Y \times \mathcal{X})$ to $\mathbb{R}_+$ such that as a function of $B \in \mathcal{B}_Y$ it is a finite measure on $(\mathcal{Y}, \mathcal{B}_Y)$ and as a function of $x \in \mathcal{X}$ it is α-integrable. Assume:*

1. *Law of Marginal, Q^X: $\mathbf{P}_X$ is a random probability measure on $(\mathcal{X}, \mathcal{B}_X)$ where $\mathbf{P}_X \sim DP(\alpha)$.*
2. *Law of Conditionals, $Q_x^{Y|X}$: $\forall x \in \mathcal{X}$, $\mathbf{P}_{Y|X}(\cdot|x)$ is a random probability measure on $(\mathcal{Y}, \mathcal{B}_Y)$ where $\mathbf{P}_{Y|X}(\cdot|x) \sim DP(\mu(\cdot, x))$.*
3. *Joint Law of Conditionals, $Q^{Y|X} = \prod_{x \in \mathcal{X}} Q_x^{Y|X}$: $\mathbf{P}_{Y|X}(\cdot|x), x \in \mathcal{X}$ are independent among themselves.*

4. *Joint Law of Marginal and Conditionals,* $Q = Q^X \times Q^{Y|X}$*:* $\mathbf{P}_X$ *is independent of* $\left\{\mathbf{P}_{Y|X}(\cdot|x)\right\}_{x\in\mathcal{X}}$.

The joint law of the marginal and conditionals, Q*, induces the law,* $\tilde{Q}$*, of the stochastic process* $\{\mathbf{P}(C)\}_{C\in\mathcal{B}}$ *through the following reparametrization:*

$$\mathbf{P}(A\times B) \stackrel{d}{=} \int_A \mathbf{P}_{Y|X}(B \mid x) d\mathbf{P}_X(x), \quad \text{for any set } A\times B \in \mathcal{B}_X \times \mathcal{B}_Y. \tag{3.7}$$

This process is called an Enriched Dirichlet process *(EDP) with parameters* α *and* μ*, and is denoted* $\mathbf{P} \sim EDP(\alpha, \mu)$.

The following theorem verifies that (3.4.1) induces a law for the random joint.

Theorem 3.4.2 *The joint law of the marginal and conditionals,* Q*, defined by the four conditions in definition (3.4.1) induces a distribution,* $\tilde{Q}$*, for the random joint probability measure.*

Proof. To prove the theorem, we must show that the map $(P_X, P_{Y|X}) \to \int_{(\cdot)} P_{Y|X}(\cdot|x) dP_X(x)$ is jointly measurable in $(P_X, P_{Y|X})$. To do so, we define a subspace of $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$ that has measure one, such that on this subspace, the mapping is measurable.

First note that in order for $\left\{\mathbf{P}_{Y|X}(\cdot|x), x \in \mathcal{X}\right\}$ to be a set of conditional random probability measures, the following two properties need to be satisfied:

1. $\forall x \in \mathcal{X}$, $\mathbf{P}_{Y|X}(\cdot|x)$ is a probability measure on $(\mathcal{Y}, \mathcal{B}_Y)$ a.s $Q_x^{Y|X}$.
2. $\forall B \in \mathcal{B}_Y$, as a function of x, $\mathbf{P}_{Y|X}(B|x)$ is $\mathcal{B}_X$ measurable a.s $Q^{Y|X}$.

The first item is satisfied since $\mathbf{P}_{Y|X}(\cdot|x) \sim \mathrm{DP}(\mu(\cdot, x))$ implies $\mathbf{P}_{Y|X}(\cdot|x) \in \mathcal{M}(\mathcal{Y})$ with probability one. The second property follows from results of Ramamoorthi and Sangalli [2006]. In particular, letting Δ be the subset of $\mathcal{M}(\mathcal{Y})^{\mathcal{X}}$ such that $P_{Y|X}$ is measurable as a function of x, they show that if $\mathbf{P}_{Y|X}(\cdot|x)$ are independent among $x \in \mathcal{X}$, then the product measure,

$Q^{Y|X} = \prod_{x \in \mathcal{X}} Q_x^{Y|X}$, given by Kolmogorov's Extension Theorem, assigns outer measure one to Δ.

Let $\mathcal{M}_D(\mathcal{X})$ denote the set of discrete probability measures on the measurable space $(\mathcal{X}, \mathcal{B}_X)$. From properties of the DP, $Q^X(\mathcal{M}_D(\mathcal{X})) = 1$. Therefore, by independence of $\mathbf{P}_X$ and $\mathbf{P}_{Y|X}$, the set $\mathcal{M}_D(\mathcal{X}) \times \Delta$ has Q-measure one. Again, by results of Ramamoorthi and Sangalli [2006], on $\mathcal{M}_D(\mathcal{X}) \times \Delta$, for $A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$, the function $(P_X, P_{Y|X}) \to \int_A P_{Y|X}(B|x) dP_X(x)$ is jointly measurable in $(P_X, P_{Y|X})$. These results imply that we can define a prior, $\tilde{Q}$, on $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ induced from Q restricted to $\mathcal{M}_D(\mathcal{X}) \times \Delta$ via the map $(P_X, P_{Y|X}) \to \int_{(\cdot)} P_{Y|X}(\cdot|x) dP_X(x)$. ∎

Remark 3. Ramamoorthi and Sangalli [2006] showed that if $\mathbf{P} \sim \mathrm{DP}(\gamma P_0)$ where $\gamma \in \mathbb{R}^+$ and $P_0 \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ is non-atomic, then

1. Law of Marginal: $\mathbf{P}_X \sim \mathrm{DP}(\gamma P_{0X})$.
2. Law of Conditionals: $\forall x \in \mathcal{X}$, $\mathbf{P}_{Y|X}(\cdot|x)$ is degenerate at some $y \in \mathcal{Y}$ with probability one.
3. Joint Law of Conditionals: $\mathbf{P}_{Y|X}(\cdot|x), x \in \mathcal{X}$ are independent among themselves.
4. Joint Law of Marginal and Conditionals: $\mathbf{P}_X$ is independent of $\left\{\mathbf{P}_{Y|X}(\cdot|x)\right\}_{x \in \mathcal{X}}$.

The EDP maintains the first, third, and fourth conditions, but relaxes the constraint on the law of the conditionals.

Obviously, the map used in Definition 3.4.1 is not $1-1$. In fact, the definition of the EDP states that the four conditions hold for the joint distribution of $(\mathbf{P}_X, \mathbf{P}_{Y|X})$ for a fixed version of the conditional, and this induces a prior on the joint. However, from the induced prior on the random joint probability measure, we can obtain the joint distribution of $\mathbf{P}_X$ and $\mathbf{P}_{Y|X}$ through the mapping $\mathbf{P} \to (\mathbf{P}_X, \mathbf{P}_{Y|X})$ defined from any version of the conditional. In the next section, we show that although the

mapping is not 1-1, the joint law of $\mathbf{P}_X$ and $\mathbf{P}_{Y|X}$ defined from any version of the conditional and the induced law of the joint probability measure still satisfies the conditions in definition (3.4.1) through an extension of the enriched Pólya urn scheme to the infinite case.

3.4.1 Enriched Pólya sequence

Similar to Blackwell and MacQueen [1973], we define an Enriched Pólya sequence which extends the enriched Pólya urn scheme to the case when $\mathcal{X}$ and $\mathcal{Y}$ are complete separable metric spaces.

Definition 3.4.3 *The sequence of random vectors* $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ *taking values in* $\mathcal{X} \times \mathcal{Y}$ *is an* Enriched Pólya sequence *with parameters* α *and* μ *if:*

1. *For* $A \in \mathcal{B}_X$ *and for all* $n \geq 1$,

$$Pr(X_1 \in A) = \frac{\alpha(A)}{\alpha(\mathcal{X})},$$

$$Pr(X_{n+1} \in A \mid X_1 = x_1, ..., X_n = x_n) = \frac{\alpha(A) + \sum_{i=1}^{n} \delta_{x_i}(A)}{\alpha(\mathcal{X}) + n}.$$

2. *For* $B \in \mathcal{B}_Y$ *and for all* $n \geq 1$,

$$Pr(Y_1 \in B \mid X_1 = x) = \frac{\mu(B, x)}{\mu(\mathcal{Y}, x)},$$

$$Pr(Y_{n+1} \in B \mid Y_1 = y_1, ..., Y_n = y_n, X_1 = x_1, ..., X_n = x_n, X_{n+1} = x) = \frac{\mu(B, x) + \sum_{j=1}^{n_x} \delta_{y_{x,j}}(B)}{\mu(\mathcal{Y}, x) + n_x},$$

where $n_x = \sum_{i=1}^{n} \mathbf{1}(x_i = x)$ *and* $\{y_{x,j}\}_{j=1}^{n_x} = \{y_i : x_i = x, i = 1, ..., n\}$.

In words, the predictive distributions characterizing the Enriched Pólya sequence can be interpreted in terms of draws from urns as follows; initially, there is an X-urn containing $\alpha(\mathcal{X})$ balls of color 0. A ball is first drawn from the X-urn, and once drawn, its true color, x_1, is revealed

(where x_1 is the realization of a draw from $P_{0X}(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathcal{X})}$). A ball of color x_1 is added to the urn along with a ball of color 0, so that the urn is now composed of $\alpha(\mathcal{X})$ balls of color 0 and one ball of color x_1. Once the true color x_1 of the X-ball is revealed, a $Y|x_1$-urn is created with $\mu(\mathcal{Y}, x_1)$ balls of color 0. Next, a ball is drawn from the $Y|x_1$-urn, and similarly, once drawn its true color is revealed to be y_1 (where y_1 is the realization of a draw from $P_{0Y|X}(\cdot|x_1) = \frac{\mu(\cdot,x_1)}{\mu(\mathcal{Y},x_1)}$). This ball is then added to the $Y|x_1$-urn along with a ball of color 0, so that the urn contains $\mu(\mathcal{Y}, x_1)$ balls of color 0 and one ball of color y_1.

At the next stage, we again first draw a ball from the X-urn. We can either draw a 0 ball or an x_1 ball. If an x_1 ball is drawn, we replace it along with another ball of the same color and then draw a Y-ball from the $Y|x_1$ urn. If the X-ball drawn is of color 0, then once drawn its true color is revealed, x_2. We add a ball of color x_2 to the X-urn and create a $Y|x_2$ urn with $\mu(\mathcal{Y}, x_2)$ balls of color 0. This process is repeated, so that a new $Y|x$ urn is created for each new value of X that is observed.

Note that if $\mathbf{P} \sim \text{EDP}(\alpha, \mu)$ and the random vectors $(X_1, Y_1), ...(X_n, Y_n)$ given $\mathbf{P} = P$ are i.i.d and distributed according to P, then $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an enriched Pólya sequence. Conversely, the following theorem proves that if $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an Enriched Pólya sequence, then given a random probability measure $\mathbf{P} = P$, the random vectors $(X_1, Y_1), ...(X_n, Y_n)$ are i.i.d and distributed according to P where the joint distribution of $(\mathbf{P}_X, \mathbf{P}_{Y|X})$ defined from any fixed version of the conditional satisfies the four conditions in definition (3.4.1). Therefore, in addition to the fact that the de Finetti measure of an Enriched Pólya sequence is an Enriched Dirichlet process, this theorem also shows that the induced law of the random joint from the four conditions in definition (3.4.1) still maintains those properties even though the mapping is not $1 - 1$.

Theorem 3.4.4 *If $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an Enriched Pólya sequence with parameters α and μ, then $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an exchangeable sequence and its de Finetti measure is an Enriched Dirichlet process with parameters (α, μ).*

Proof. For a quick sketch of the proof, we start by showing that the sequence $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is exchangeable, and then apply de Finetti's Theorem. Next, after reparametrizing in terms of the marginal and conditionals, we verify the de Finetti measure satisfies the four conditions in the definition of the EDP.

First, note that the sequence $\{X_n\}_{n\in\mathbb{N}}$ is a Pólya sequence with parameter α. Recall that the predictive distribution of a Pólya sequence converges to a discrete random probability measure with positive mass at the countable number of unique values of the sequence almost surely with respect to the exchangeable law. Therefore, given $X_1 = x_1, ..., X_n = x_n$ and letting $U(x_1, ..., x_n)$ denote the set of the unique values of $\{x_1, ..., x_n\}$, we have that for $x^* \in U(x_1, ..., x_n)$,

$$n_{x^*} = \sum_{i=1}^{n} \mathbf{1}(x^* = x_i) \to \infty \text{ as } n \to \infty,$$

almost surely with respect to the exchangeable law. This implies that given $\{X_n = x_n\}_{n\in\mathbb{N}}$, for any $x^* \in U(\{x_n\}_{n\in\mathbb{N}})$, the set of random variables,

$$\{Y_{x^*,j}\} = \{Y_i : X_i = x^*, i \in \mathbb{N} | \{X_n = x_n\}_{n\in\mathbb{N}}\}$$

is a countable sequence. Furthermore, by assumption, for $x_1^* \neq x_2^* \in U(\{x_n\}_{n\in\mathbb{N}})$, the sequences $\{Y_{x_1^*,j}\}_{j\in\mathbb{N}}$ and $\{Y_{x_2^*,j}\}_{j\in\mathbb{N}}$ are independent Pólya sequences with parameters $\mu(\cdot, x_1^*)$ and $\mu(\cdot, x_2^*)$ respectively. These observations imply exchangeability of the sequence $\{X_n, Y_n\}_{n\in\mathbb{N}}$, as shown in the following argument.

$$\begin{aligned} &Pr(X_1 \in A_1, Y_1 \in B_1, ..., X_n \in A_n, Y_n \in B_n) \qquad (3.8)\\ &\quad = \int_{\times_{h=1}^{n} A_h} Pr(Y_1 \in B_1, ..., Y_n \in B_n | x_1, \ldots, x_n) dPr(x_1, \ldots, x_n). \end{aligned}$$

By independence of $\{Y_{x_1^*,j}\}_{j=1}^{n_{x_1^*}}$ and $\{Y_{x_2^*,j}\}_{j=1}^{n_{x_2^*}}$ for $x_1^* \neq x_2^* \in U(x_1, ..., x_n)$, we have that (3.8) is equal to:

$$\int_{\times_{h=1}^{n} A_h} \prod_{x^*\in U(x_1,\ldots,x_n)} Pr(Y_{x^*,1} \in B_{x^*,1}, .., Y_{x^*,n_{x^*}} \in B_{x^*,n_{x^*}}) dPr(x_1, \ldots, x_n). \qquad (3.9)$$

A permutation, π, of the sets $(x_1 \times B_1), ..., (x_n \times B_n)$, is equivalent to the same permutation, π, of $(x_1, ..., x_n)$ and for $x^* \in U(x_{\pi(1)}, ..., x_{\pi(n)})$, a permutation, γ_{x^*}, of $(B_{x^*,1}, ..., B_{x^*,n_{x^*}})$. To keep notation concise, we will let $U_{\pi,n}$ represent $U(x_{\pi(1)}, ..., x_{\pi(n)})$ (and similarly, U_n represent $U(x_1, ..., x_n)$).The term inside the integral is invariant to the permutation, π, of $(x_1, ..., x_n)$, and due to exchangeability of Pólya sequences, the laws of the random vectors $\{X_i\}_{i=1}^n$ and $\{Y_{x^*,j}\}_{j=1}^{n_x^*}$ are invariant to the permutations π and γ_{x^*} respectively. Thus, (3.9) is equal to:

$$\begin{aligned}
&\int\limits_{\times_{h=1}^n A_{\pi(h)}} \prod_{x^* \in U_{\pi,n}} Pr(Y_{x^*,1} \in B_{\gamma_{x^*}(1)}, ..., Y_{x^*,n_{x^*}} \in B_{\gamma_{x^*}(n_{x*})}) dPr(x_{\pi(1:n)}) \\
&= \int\limits_{\times_{h=1}^n A_{\pi(h)}} Pr(Y_1 \in B_{\pi(1)}, ..., Y_n \in B_{\pi(n)} | x_{\pi(1:n)}) dPr(x_{\pi(1:n)}) \\
&= Pr(X_1 \in A_{\pi(1)}, Y_1 \in B_{\pi(1)}, ..., X_n \in A_{\pi(n)}, Y_n \in B_{\pi(n)}),
\end{aligned}$$

where $x_{\pi(1:n)} = (x_{\pi(1)}, ..., x_{\pi(n)})$.

De Finetti's Representation Theorem states that there exists a random probability measure, $\mathbf{P}$, with distribution $\tilde{Q}$ on $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ such that:

$$\begin{aligned}
&Pr(X_1 \in A_1, Y_1 \in B_1, ., X_n \in A_n, Y_n \in B_n) \\
&= \int\limits_{\mathcal{M}(\mathcal{X} \times \mathcal{Y})} \prod_{h=1}^n P(A_h \times B_h) d\tilde{Q}(P),
\end{aligned} \tag{3.10}$$

and $\frac{1}{n} \sum_{h=1}^n \delta_{A \times B}(X_h, Y_h) \xrightarrow{d} \mathbf{P}(A \times B)$ a.s. with respect to the exchangeable law as $n \to \infty$ where $\mathbf{P} \sim \tilde{Q}$. The distribution $\tilde{Q}$ determines the joint distribution, Q, of the marginal and a fixed version of the conditionals. Reparametrizing in terms of the marginal and conditionals implies:

$$\begin{aligned}
&Pr(X_1 \in A_1, Y_1 \in B_1, ., X_n \in A_n, Y_n \in B_n) \\
&= \int\limits_{\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{Y})^{\mathcal{X}}} \prod_{h=1}^n \int\limits_{A_h} P_{Y|X}(B_h|x) dP_X(x) dQ(P_X, \prod_{x \in \mathcal{X}} P_{Y|X}(\cdot|x)).
\end{aligned} \tag{3.11}$$

A simple application of the results of Blackwell and MacQueen [1973] for Pólya urn sequences, verifies that the first two conditions in the definition of the EDP hold. In particular, for any finite partition $A_1, ..., A_k \subseteq \mathcal{B}_{\mathcal{X}}$, define the simple measurable function, $\phi(x) = i$ if $x \in A_i$ for $i = 1, ..., k$. Noting that $\{\phi(X_n)\}_{n\in\mathbb{N}}$, is a Pólya sequence with parameter $\alpha \circ (\phi)^{-1}$ taking values in the finite space $\{1, ..., k\}$, implies:

$$\begin{aligned}\mathbf{P}_X(\phi^{-1}(1), ..., \mathbf{P}_X(\phi^{-1}(k))) &\sim \text{Dir}(\alpha(\phi^{-1}(1)), ..., \alpha(\phi^{-1}(k))) \\ \Leftrightarrow \mathbf{P}_X(A_1), .., \mathbf{P}_X(A_k) &\sim \text{Dir}(\alpha(A_1), .., \alpha(A_k)).\end{aligned}$$

Similarly, for any finite partition $B_1, .., B_m \subseteq \mathcal{B}_{\mathcal{Y}}$, define the simple measurable function $\varphi(y) = j$ if $y \in B_j$. For any $x^* \in U(\{x_n\}_{n\in\mathbb{N}})$, the sequence $\{\varphi(Y_{x^*,j})\}_{j\in\mathbb{N}}$ is a Pólya sequence taking values in the finite space $\{1, ..., m\}$ with parameter $\mu(\varphi^{-1}(\cdot), x^*)$. Again, it follows that:

$$\begin{aligned}\mathbf{P}_{Y|X}(\varphi^{-1}(1)|x^*), ..., \mathbf{P}_{Y|X}(\varphi^{-1}(m)|x^*) &\sim \text{Dir}(\mu(\varphi^{-1}(1), x^*), ..., \mu(\varphi^{-1}(m), x^*)) \\ \Leftrightarrow \mathbf{P}_{Y|X}(B_1|x^*), ..., \mathbf{P}_{Y|X}(B_m|x^*) &\sim \text{Dir}(\mu(B_1, x^*), ..., \mu(B_m, x^*)).\end{aligned} \tag{3.12}$$

The unique values of the Pólya sequence are actually draws from $P_{0X}(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathcal{X})}$ and can therefore take any value in $\mathcal{X}$. Thus, (3.12) holds for any $x \in \mathcal{X}$. Finally, we need to show the last two conditions in the definition of the EDP hold. Exchangeability of the pairs implies exchangeability of the sequence $\{Y_i|X_i = x_i\}_{i\in\mathbb{N}}$. Therefore, by de Finetti's theorem:

$$Pr(Y_1 \in B_1, ..., Y_n \in B_n|x_1, ..., x_n) \tag{3.13}$$

$$= \int_{\mathcal{P}(\mathcal{B}_Y)^{U_n}} \prod_{x^*\in U_n} \prod_{j=1}^{n_{x^*}} P_{Y|X}(B_{x^*,j}|x^*) dQ_{U_n}^{Y|X}(\prod_{x^*\in U_n} P_{Y|X}(\cdot|x^*)). \tag{3.14}$$

Independence of the exchangeable sequences $\left\{Y_{x_1^*,j}\right\}_{j\in\mathbb{N}}$ and $\left\{Y_{x_2^*,j}\right\}_{j\in\mathbb{N}}$ for $x_1^* \neq x_2^*$ implies:

$$\begin{aligned}
&Pr(Y_1 \in B_1, ..., Y_n \in B_n | x_1, ..., x_n) \\
&= \prod_{x^* \in U_n} Pr(Y_{x^*,1} \in B_{x^*,1}, ..., Y_{x^*,n_{x^*}} \in B_{x^*,n_{x^*}}) \\
&= \prod_{x^* \in U_n} \int_{\mathcal{P}(\mathcal{B}_Y)} \prod_{j=1}^{n_{x^*}} P_{Y|X}(B_{x^*,j}|x^*) dQ_{x^*}^{Y|X}(P_{Y|X}(\cdot|x^*)). \qquad (3.15)
\end{aligned}$$

Comparing (3.14) and (3.15) shows that $Q_{U_n}^{Y|X} = \prod_{x^* \in U_n} Q_{x^*}^{Y|X}$. Since the unique values of $\{x_1, ...x_n\}$ are realizations of P_{0X} and can take any value in $\mathcal{X}$, independence of $\left\{\mathbf{P}_{Y|X}(\cdot|x)\right\}_{x\in\mathcal{X}}$ among $x \in \mathcal{X}$ follows. Therefore, (3.13) can be equivalently written as:

$$\begin{aligned}
&Pr(Y_1 \in B_1, ..., Y_n \in B_n | x_1, ..., x_n) \\
&= \int_{\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}} \prod_{h=1}^{n} P_{Y|X}(B_h|x_h) d(\prod_{x\in\mathcal{X}} Q_x^{Y|X}(P_{Y|X}(\cdot|x))).
\end{aligned}$$

Now combining this result with the fact that $\{X_n\}_{n\in\mathbb{N}}$ is an exchangeable sequence implies:

$$\begin{aligned}
&Pr(X_1 \in A_1, Y_1 \in B_1, ., X_n \in A_n, Y_n \in B_n) \\
&= \int_{\times_{h=1}^n A_h} Pr(Y_1 \in B_1, ., Y_n \in B_n | x_1, ., x_n) dPr(x_1, ., x_n) \\
&= \int_{\mathcal{M}(\mathcal{X})} \int_{\times_{h=1}^n A_h} \int_{\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}} \prod_{h=1}^{n} P_{Y|X}(B_h|x_h) \\
&\qquad d(\prod_{x\in\mathcal{X}} Q_x^{Y|X}(P_{Y|X}(\cdot|x))) d(\prod_{h=1}^{n} P_X(x_h)) dQ^X(P_X) \\
&= \int_{\mathcal{M}(\mathcal{X})} \int_{\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}} \prod_{h=1}^{n} \int_{A_h} P_{Y|X}(B_h \mid x_h) dP_X(x_h) \\
&\qquad d(\prod_{x\in\mathcal{X}} Q_x^{Y|X}(P_{Y|X}(\cdot \mid x))) dQ^X(P_X). \qquad (3.16)
\end{aligned}$$

Comparing (3.11) with (3.16) implies that $Q = Q^X \times \prod_{x\in\mathcal{X}} Q_x^{Y|X}$, i.e independence of $\mathbf{P}_X$ and $\left\{\mathbf{P}_{Y|X}(\cdot|x)\right\}_{x\in\mathcal{X}}$. ∎

3.4.2 Properties

Define $P_{0X}(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathcal{X})}$ and for every $x \in \mathcal{X}$, $P_{0Y|X}(\cdot|x) = \frac{\mu(\cdot,x)}{\mu(\mathcal{Y},x)}$. From well-known properties of the Dirichlet distribution, we have:

Proposition 3.4.5 *If* $\mathbf{P} \sim EDP(\alpha,\mu)$*, for* $A \in \mathcal{B}_X, B \in \mathcal{B}_Y$*,*

$$\begin{aligned}
E[\mathbf{P}_X(A)] &= P_{0X}(A),\\
Var(\mathbf{P}_X(A)) &= \frac{P_{0X}(A)(1-P_{0X}(A))}{\alpha(\mathcal{X})+1};\\
E[\mathbf{P}_{Y|X}(B \mid x)] &= P_{0Y|X}(B|x) \quad \forall x \in \mathcal{X},\\
Var(\mathbf{P}_{Y|X}(B|x)) &= \frac{P_{0Y|X}(B \mid x)(1-P_{0Y|X}(B|x))}{\mu(\mathcal{Y},x)+1} \quad \forall x \in \mathcal{X};\\
E[\mathbf{P}(A\times B)] &= \int_A P_{0Y|X}(B|x)dP_{0X}(x) := P_0(A\times B).
\end{aligned}$$

Therefore, similar to the DP, the location of the EDP is determined by the base measure P_0, but the there are now many more parameters to control the precision, namely $\alpha(\mathcal{X})$ and $\mu(\mathcal{Y},x)$ for every $x \in \mathcal{X}$. The parameters of the EDP may equivalently be parametrized in terms of the base measure P_0 and the precision parameter $\alpha(\mathcal{X})$ of the marginal and the collection of precision parameters $\mu(\mathcal{Y},x)$ for the conditionals.

The following proposition states that the DP is in fact a special case of the EDP.

Proposition 3.4.6 $\mathbf{P} \sim EDP(\alpha,\mu)$ *with* $\mu(\mathcal{Y},x) = \alpha(\{x\}), \forall x \in \mathcal{X}$ *is equivalent to* $\mathbf{P} \sim DP(\alpha(\mathcal{X})P_0)$*.*

Proof. The proof relies on the urn characterization of both processes; we show that an Enriched Pólya sequence is equivalent to a Pólya sequence with parameter $\alpha(\mathcal{X})\,P_0(\cdot)$, if $\mu(\mathcal{Y},x) = \alpha(\{x\})$, $\forall x \in \mathcal{X}$. For an Enriched

Pólya sequence with parameters α, μ and for $A \in \mathcal{B}_X, B \in \mathcal{B}_Y$, since

$$\lim_{\mu(\mathcal{Y},x)\to\alpha(\{x\})} Pr(Y_1 \in B \mid X_1 = x) = P_{0Y|X}(B|x),$$

then if $\mu(\mathcal{Y}, x) = \alpha(\{x\})$, $\forall x \in \mathcal{X}$,

$$Pr(X_1 \in A, Y_1 \in B) = P_0(A \times B).$$

The joint predictive distribution is given by

$$\begin{aligned} &Pr(X_{n+1} \in A, Y_{n+1} \in B | X_1 = x_1, Y_1 = y_1, ..., X_n = x_n, Y_n = y_n) \\ &= \int_A \frac{\mu(B,x) + \sum_{j=1}^{n_x} \delta_{y_{x,j}}(B)}{\mu(\mathcal{Y},x) + n_x} d\left(\frac{\alpha + \sum_{i=1}^{n} \delta_{x_i}}{\alpha(\mathcal{X}) + n}\right)(x). \end{aligned} \tag{3.17}$$

Rewriting this as the sum of the integrals over the sets $A \setminus \{x_1, ..., x_n\}$ and $A \cap \{x_1, ..., x_n\}$ and replacing $\mu(\mathcal{Y}, x)$ with $\alpha(\{x\})$, we get that (3.17) is equal to

$$\begin{aligned} &\frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X}) + n} P_0(A \setminus \{x_1, ..., x_n\} \times B) \\ &+ \sum_{x \in A \cap \{x_1,...,x_n\}} \frac{\alpha(\{x\}) P_{0Y|X}(B|x) + \sum_{j=1}^{n_x} \delta_{y_{x,j}}(B)}{\alpha(\{x\}) + n_x} \frac{\alpha(\{x\}) + n_x}{\alpha(\mathcal{X}) + n} \\ &= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X}) + n} P_0(A \times B) + \frac{n}{\alpha(\mathcal{X}) + n} \sum_{i=1}^{n} \frac{\delta_{x_i,y_i}(A,B)}{n}. \end{aligned}$$

∎

As a by-product of this proposition, if $\mathbf{P} \sim \mathrm{DP}(\gamma P_0)$, the law of the random conditionals is $\mathbf{P}_{Y|X}(\cdot|x) \sim \mathrm{DP}(\gamma P_{0X}(\{x\}) P_{0Y|X}(\cdot|x))$, where $\mathbf{P}_{Y|X}(\cdot|x)$ are independent among $x \in \mathcal{X}$. In general, the marginal base measure P_{0X} can assign positive mass to countably many locations. Any random conditional probability measure associated with x that has positive mass under the marginal base measure will be a DP with precision parameter equivalent to the mass of x under the marginal base measure times γ. Since a DP with precision parameter 0 is degenerate at a random location with probability one, the random conditional probability

measures associated with all other x's will be degenerate at some $y \in \mathcal{Y}$ with probability one. Thus, in the case when P_0 is non-atomic, a DP implies assuming the conditionals are independent and degenerate a.s., which is consistent with results in Ramamoorthi and Sangalli [2006] given in *Remark 3*. The EDP relaxes the constraint required by the DP that the precision parameters of the conditionals are $\gamma P_{0X}(\{x\})$, allowing more flexibility.

As noted by Ferguson [1973], a prior for nonparametric problems should have large topological support. The following theorem shows that the EDP has full weak support. Here, $\mathcal{X} = \mathbb{R}^{p_1}$ and $\mathcal{Y} = \mathbb{R}^{p_2}$, implying $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^p$ where $p = p_1 + p_2$.

Theorem 3.4.7 *Let S_0 denote the topological support of P_0. If $\mathbf{P} \sim EDP(\alpha, \mu)$, then the topological support of $\mathbf{P}$ is*

$$M_0 = \{P \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) : \textit{topological support}(P) \subseteq S_0\}.$$

Proof. This proof is based on the proof of Theorem 3.2.4 in Ghosh and Ramamoorthi [2003]. To show M_0 is the topological support - the smallest closed set of measure one - it is enough to show that M_0 is a closed set of measure one, such that for every $\Pi \in M_0$, $Q(U) > 0$ for any neighborhood U of Π.

First, we show M_0 is closed. If $P_n \in M_0$, then $P_n(S_0) = 1$ for all n and if $P_n \overset{weakly}{\to} P$, then for any closed set $C \in \mathcal{B}$, $\limsup_n P_n(C) \leq P(C)$. Together these imply $P(S_0) = 1$, or equivalently, $P \in M_0$.

Secondly, the set M_0 has measure one. This follows from the square breaking construction of $\mathbf{P}$ (see Proposition 3.4.11). Since $X_i^*, Y_{j|i}^* \sim P_0$ implies $\delta_{\tilde{X}_i, \tilde{Y}_{j|i}^*}(S_0) = 1$ a.s., $\sum_{i=1}^{\infty} w_i = 1$ a.s., and for all i, $\sum_{j=1}^{\infty} w_{j|i} = 1$ a.s, then $\mathbf{P}(S_0) = 1$ a.s. ($\Leftrightarrow Q(M_0) = 1$).

Lastly, our theorem will be proved if we show that for any $\Pi \in M_0$ and any neighborhood U of Π, $Q(U) > 0$. By extension of Proposition 2.5.2 in Ghosh and Ramamoorthi [2003], there exists points $q_{1,j} < ... < q_{n_j,j}$ in

$\mathbb{R}$ for $j = 1, .., p$, and $\delta > 0$, such that

$$U^* = \left\{ P \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) : |P(\prod_{j=1}^{p}[q_{i_j,j}, q_{i_j+1,j})) - \Pi(\prod_{j=1}^{p}[q_{i_j,j}, q_{i_j+1,j}))| < \delta \right.$$
$$\left. \text{and } \Pi(\partial \prod_{j=1}^{p}[q_{i_j,j}, q_{i_j+1,j})) = 0 \text{ for } i = 1, ..., n_j, j = 1, ..., p \right\} \subseteq U.$$

Define $A_{i_1,..,i_{p_1}} = \prod_{j=1}^{p_1}[q_{i_j,j}, q_{i_j+1,j})$ and $B_{i_{p_1+1},..,i_p} = \prod_{j=p_1+1}^{p}[q_{i_j,j}, q_{i_j+1,j})$ and without loss of generality, we denote these sets as $A_1, ..., A_N$ and $B_1, ..., B_M$. If $P_0(A_n \times B_m) = 0$, then $\delta_{\tilde{X}_i,\tilde{Y}_{j|i}}(S_0) = 0$ a.s. and $\mathbf{P}(A_n \times B_m)$ is degenerate 0. In addition, $P_0(A_n \times B_m) = 0$ combined with the facts that $\Pi(\partial A_n \times B_m) = 0$ and $\Pi(S_0) = 1$, imply that $\Pi(A_n \times B_n) = 0$. Therefore, $|\mathbf{P}(A_n \times B_m) - \Pi(A_n \times B_m)| = 0$ a.s.. If $P_0(A_n \times B_m) > 0$, then $\delta_{\tilde{X}_i,\tilde{Y}_{j|i}}(A_n \times B_m) = 1$ with positive probability. Thus, the square breaking construction implies that $Q(U^*) > 0$. ∎

3.4.3 Posterior

Just as the finite dimensional Enriched Dirichlet distribution is conjugate to the multinomial likelihood, the Enriched Dirichlet process is also conjugate for estimating an unknown distribution from exchangeable data. More precisely,

Proposition 3.4.8 *If* $(X_i, Y_i) \mid \mathbf{P} = P \overset{iid}{\sim} P$, *where* $\mathbf{P} \sim EDP(\alpha, \mu)$, *then*

$$\mathbf{P} \mid x_1, y_1, ..., x_n, y_n \sim EDP(\alpha_n, \mu_n),$$

where

$$\alpha_n = \alpha + \sum_{i=1}^{n} \delta_{x_i},$$

and for all $x \in \mathcal{X}$,

$$\mu_n(\cdot, x) = \mu(\cdot, x) + \sum_{j=1}^{n_x} \delta_{y_{x,j}},$$

with $\quad n_x = \sum_{i=1}^{n} \mathbf{1}(x_i = x)$ *and* $\{y_{x,j}\}_{j=1}^{n_x} = \{y_j : x_j = x\}$.

The proof of conjugacy is straightforward; one simply has to demonstrate that given the random sample the four conditions in the definition of EDP hold with the updated parameters specified above. The first two conditions, the fact that the marginal and conditionals are DPs with updated parameters, follow from conjugacy of the DP. The last two conditions, independence of the marginal and conditionals and independence among the conditionals, follow by combining the fact that a priori independence holds with independence of the random vectors $(X_1, ..., X_n)$ and $(Y_1, ..., Y_n | X_1 = x_1, ..., X_n = x_n)$ and independence of the random vectors $\{Y_{x,j}\}_{j=1}^{n_x}$ among $x \in \mathcal{X}$.

Posterior consistency is a frequentist validation tool that is useful in Bayesian nonparametric inference where the infinite dimension of the parameter space can make specification of a prior challenging and cause the prior to strongly influence the posterior even with large amounts of data. One of the reasons that makes the Dirichlet process so appealing is that the posterior is weakly consistent for any probability measure, Π, on the product space under the assumption that the sequence of random vectors are distributed according to the i.i.d. product measure Π^∞. Another important property that the EDP maintains is posterior consistency. The proof requires that for a set $A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$, the posterior expectation of $\mathbf{P}(A \times B)$ converges to $\Pi(A \times B)$ a.s. Π^∞ and its posterior variance goes to zero. In the following lemma, the variance of the probability over a set $A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$ is specified.

Lemma 3.4.9 *If* $\mathbf{P} \sim EDP(\alpha, \mu)$*, for* $A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$*,*

$$Var(\mathbf{P}(A \times B)) = \frac{1}{\alpha(\mathcal{X})+1} \int_A \frac{P_{0Y|X}(B|x)(1+\mu(\mathcal{Y},x)P_{0Y|X}(B|x))}{\mu(\mathcal{Y},x)+1} dP_{0X}(x) \tag{I_1}$$

$$+ \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+1} \int_A \int_{\{x\}} \frac{P_{0Y|X}(B|x)(1-P_{0Y|X}(B|x))}{\mu(\mathcal{Y},x)+1} dP_{0X}(x') dP_{0X}(x) \tag{I_2}$$

$$- \frac{1}{\alpha(\mathcal{X})+1} \int_A \int_{\{x\}} P_{0Y|X}(B|x)^2 dP_{0X}(x') dP_{0X}(x) \tag{I_3}$$

$$- \frac{1}{\alpha(\mathcal{X})+1} \int_A \int_{A\setminus\{x\}} P_{0Y|X}(B|x') P_{0Y|X}(B|x) dP_{0X}(x') dP_{0X}(x). \tag{I_4}$$

Proof.

$$\mathrm{E}[\mathbf{P}(A \times B)^2] = \mathrm{E}[\sum_{i=1}^{\infty} w_i^2 \mathbf{P}_{Y|X}(B|\tilde{X}_i)^2 \delta_{\tilde{X}_i}(A)] \tag{J_1}$$

$$+ \mathrm{E}[\sum_{i=1}^{\infty} \sum_{j\neq i} w_i w_j \mathbf{P}_{Y|X}(B|\tilde{X}_i)^2 \delta_{\tilde{X}_i}(A) \delta_{\tilde{X}_j}(\{\tilde{X}_i\})] \tag{J_2}$$

$$+ \mathrm{E}[\sum_{i=1}^{\infty} \sum_{j\neq i} w_i w_j \mathbf{P}_{Y|X}(B|\tilde{X}_i) \mathbf{P}_{Y|X}(B|\tilde{X}_j) \delta_{\tilde{X}_i}(A) \delta_{\tilde{X}_j}(A \setminus \{\tilde{X}_i\})]. \tag{J_3}$$

Using the fact that $E_w[\sum_{i=1}^{\infty} w_i^2] = \frac{1}{\alpha(\mathcal{X})+1}$ and properties of the Dirichlet distribution,

$$(J_1) = \mathrm{E}_w[\sum_{i=1}^{\infty} w_i^2 \mathrm{E}_{\tilde{X}}[\mathrm{E}_{Q^{Y|X}}[\mathbf{P}_{Y|X}(B|\tilde{X}_i)^2|\tilde{X}_i] \delta_{\tilde{X}_i}(A)]]$$

$$= \frac{1}{\alpha(\mathcal{X})+1} \int_A \frac{P_{0Y|X}(B|x)(1+\mu(\mathcal{Y},x)P_{0Y|X}(B|x))}{\mu(\mathcal{Y},x)+1} dP_{0X}(x).$$

Now, using the fact that $E_w[\sum_{i=1}^{\infty}\sum_{i\neq j} w_i w_j] = \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+1}$ and, again, properties of the Dirichlet distribution,

$$
\begin{aligned}
(J_2) &= \mathrm{E}_w[\sum_{i=1}^{\infty}\sum_{i\neq j} w_i w_j \mathrm{E}_{\tilde{X}}[\mathrm{E}_{Q^{Y|X}}[\mathbf{P}_{Y|X}(B|\tilde{X}_i)^2|\tilde{X}_i]\delta_{\tilde{X}_i}(A)\delta_{\tilde{X}_j}(\{\tilde{X}_i\})]] \\
&= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+1}\int_A\int_{\{x\}} \frac{P_{0Y|X}(B|x)(1+\mu(\mathcal{Y},x)P_{0Y|X}(B|x))}{\mu(\mathcal{Y},x)+1} dP_{0X}(x')dP_{0X}(x), \\
(J_3) &= \mathrm{E}_w[\sum_{i=1}^{\infty}\sum_{i\neq j} w_i w_j \\
&\qquad \mathrm{E}_{\tilde{X}}[\mathrm{E}_{Q^{Y|X}}[\mathbf{P}_{Y|X}(B|\tilde{X}_i)\mathbf{P}_{Y|X}(B|\tilde{X}_j)|\tilde{X}_i,\tilde{X}_j]\delta_{\tilde{X}_i}(A)\delta_{\tilde{X}_j}(A\setminus\{\tilde{X}_i\})]] \\
&= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+1}\int_A\int_{A\setminus\{x\}} P_{0Y|X}(B|x')P_{0Y|X}(B|x)dP_{0X}(x')dP_{0X}(x).
\end{aligned}
$$

The result is obtained following some algebra. ∎

Theorem 3.4.10 *If* $\mathbf{P} \sim EDP(\alpha,\mu)$*, then, for* $\Pi \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})$*, the posterior distribution,* Q_n*, of* $\mathbf{P}$ *converges weakly to* δ_Π *for* $n\to\infty$*, a.s.* Π^∞*.*

Proof. First, we show that $E[\mathbf{P}(A\times B)|X_1 = x_1, Y_1 = y_1, ..., X_n = x_n, Y_n = y_n] \to \Pi(A\times B)$ a.s. Π^∞.

$$
\begin{aligned}
&E[\mathbf{P}(A\times B)|X_1 = x_1, Y_1 = y_1, ..., X_n = x_n, Y_n = y_n] \\
&= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n} P_0(A\setminus\{x_1,...,x_n\}\times B) \\
&\qquad + \sum_{x\in A\cap\{x_1,...,x_n\}} \frac{\mu(\mathcal{Y},x)+\sum_{j=1}^{n_x}\delta_{y_{x,j}}(B)}{\alpha(\mathcal{X})+n}\frac{\alpha(x)+n_x}{\mu(\mathcal{Y},x)+n_x} \\
&\sim \frac{1}{n}\sum_{x\in A\cap\{x_1,...,x_n\}}\sum_{j=1}^{n_x}\delta_{y_{x,j}}(B) = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i,y_i}(A,B) \\
&\to \Pi(A\times B) \text{ a.s } \Pi^\infty.
\end{aligned}
$$

Using lemma (3.4.9), we show the posterior variance of $\mathbf{P}(A\times B)$ goes to 0, by showing each of the four terms in (3.4.9) goes to 0. Since

$$\frac{\alpha_n(A)}{\alpha_n(\mathcal{X})} \sim \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(A),$$

and for $x \in \{x_1, ..., x_n\}$,

$$\frac{\mu_n(B, x)}{\mu_n(\mathcal{Y}, x)} \sim \frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B),$$

we have that

$$\begin{aligned}
(I_1) &\sim \frac{1}{n} \int_A (\frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B))(\frac{1}{n_x} + \frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B)) d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x)) \\
&\to 0, \\
(I_2) &\sim \int_A \int_{\{x\}} \frac{1}{n_x} (\frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B))(\frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B^c)) \\
&\qquad d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x')) d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x)) \\
&\to 0, \\
(I_3) &\sim -\frac{1}{n} \int_A \int_{\{x\}} (\frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B))^2 d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x')) d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x)) \\
&\to 0, \\
(I_4) &\sim -\frac{1}{n} \int_A \int_{A \setminus \{x\}} (\frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B))(\frac{1}{n_{x'}} \sum_{i=1}^{n_{x'}} \delta_{y_{x',j}}(B)) \\
&\qquad d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x')) d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x)) \\
&\to 0.
\end{aligned}$$

This holds for any finite collection of sets. By a straightforward extension of Theorem 2.5.2 of Ghosh and Ramamoorthi [2003], this implies weak convergence of Q_n to δ_Π a.s. Π^∞. ∎

3.4.4 Square-breaking construction

The following square-breaking representation of the EDP is a direct result of Sethuraman's stick-breaking representation of the DP (Sethuraman [1994]).

Proposition 3.4.11 *If* $\mathbf{P} \sim EDP(\alpha, \mu)$*, it has the following square-breaking a.s. representation*

$$\mathbf{P} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} w_i w_{j|i} \delta_{\tilde{X}_i, \tilde{Y}_{j|i}},$$

where $w_1 = v_1$ *and* $w_i = v_i \prod_{i'=1}^{i-1}(1 - v_{i'})$ *for* $i > 1$*, with*

$$v_i \overset{iid}{\sim} Beta(1, \alpha(\mathcal{X})),$$

$$\tilde{X}_i \overset{iid}{\sim} P_{0X},$$

and for $i = 1, 2, ...,$ $w_{1|i} = v_{1|i}$ *and* $w_{j|i} = v_{j|i} \prod_{j'=1}^{j-1}(1 - v_{j'|i})$ *for* $j > 1$*, with*

$$v_{j|i} | \tilde{X}_i = \tilde{x}_i \overset{ind}{\sim} Beta(1, \mu(\mathcal{Y}, \tilde{x}_i)),$$

$$\tilde{Y}_{j|i} | \tilde{X}_i = \tilde{x}_i \overset{ind}{\sim} P_{0Y|X}(\cdot | \tilde{x}_i),$$

and the sequences $\{v_i\}_{i=1}^{\infty}$*;* $\{\tilde{X}_i\}_{i=1}^{\infty}$*;* $\{v_{j|1} | \tilde{X}_1 = \tilde{x}_1\}_{j=1}^{\infty}, \{v_{j|2} | \tilde{X}_2 = \tilde{x}_2\}_{j=1}^{\infty}, ...;$ *and* $\{\tilde{Y}_{j|1} | \tilde{X}_1 = \tilde{x}_1\}_{j=1}^{\infty}, \{\tilde{Y}_{j|2} | \tilde{X}_2 = \tilde{x}_2\}_{j=1}^{\infty}, ...$ *are independent.*

For an interpretation of this proposition, consider a square of area one; we break off rectangles of the square defined by a width of w_i and length of $w_{j|i}$ and we assign the area of that rectangle, $w_i w_{j|i}$, to a random location $(\tilde{X}_i, \tilde{Y}_{j|i})$.

Note that while a closed form for the finite dimensional distributions of $\mathbf{P}_Y$ may not be available, we can obtain a square-breaking construction for the random marginal probability measure on $(\mathcal{Y}, \mathcal{B}_Y)$,

$$\mathbf{P}_Y = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} w_i w_{j|i} \delta_{\tilde{Y}_{j|i}},$$

where the distribution of $\{w_i\}$, $\{w_{j|i}\}$, $\{\tilde{Y}_{j|i}\}$ is specified above.

3.4.5 Clustering structure

The clustering structure in a sample from $\mathbf{P} \sim \text{EDP}$ is characterized by the predictive rule. In particular, the predictive rule states that if P_0 is non-atomic, for $A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$:

$$Pr(X_{n+1} \in A, Y_{n+1} \in B | x_1, y_1, ..., x_n, y_n)$$
$$= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X}) + n} P_0(A \times B) + \sum_{x_i^* \in A} \frac{n_i}{\alpha(\mathcal{X}) + n} \left(\frac{\mu(B, x_i^*) + \sum_{j=1}^{n_i} \delta_{y_{x_i^*,j}}(B)}{\mu(\mathcal{Y}, x_i^*) + n_i} \right),$$

where $(x_1^*, ..., x_k^*)$ denotes the unique values of $(x_1, ..., x_n)$, k is the number of unique values, and $n_i = \sum_{i'=1}^{n} \mathbf{1}(x_{i'} = x_i^*)$. Thus, the pair (X_{n+1}, Y_{n+1}) is either a "new-new" , "old-new" , or "old-old" pair with probabilities obtained by replacing the set $A \times B$ with the sets $(\mathcal{X} \setminus \{x_1, ..., x_n\}) \times (\mathcal{Y} \setminus \{y_1, ..., y_n\})$, $\{x_1, ..., x_n\} \times (\mathcal{Y} \setminus \{y_1, ..., y_n\})$, or $\{x_1, ..., x_n\} \times \{y_1, ..., y_n\}$ respectively. Let $(y_{1|i}^*, ..., y_{k_i|i}^*)$ be the unique values of $(y_{x_i^*,1}, ..., y_{x_i^*,n_i})$ where k_i is the number of unique values in this set and $n_{i,j} = \sum_{j'=1}^{n_i} \mathbf{1}(y_{x_i^*,j'} = y_{j|i}^*)$. Succinctly, the clustering structure is described as follows:

$$X_{n+1}, Y_{n+1} | x_{1:n}, y_{1:n} = \begin{cases} (x_{k+1}^*, y_{1|k+1}^*) & \text{wp } \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n}, \\ (x_i^*, y_{k_i+1|i}^*) & \text{wp } \frac{n_i}{\alpha(\mathcal{X})+n} \frac{\mu(\mathcal{Y}, x_i^*)}{\mu(\mathcal{Y}, x_i^*)+n_i}, \\ (x_i^*, y_{j|i}^*) & \text{wp } \frac{n_i}{\alpha(\mathcal{X})+n} \frac{n_{i,j}}{\mu(\mathcal{Y}, x_i^*)+n_i}, \end{cases}$$

where $(X_{k+1}^*, Y_{1|k+1}^*) \sim P_0$ and $Y_{k_i+1|i}^* \sim P_0(\cdot | x_i^*)$. This gives a "two-level" clustering which reduces to the global clustering of the DP if $\mu(\mathcal{Y}, x) = 0$ for all $x \in \mathcal{X}$.

The availability of an analytically computable urn scheme is a particularly attractive feature of the EDP over other extensions of the DP, such as Dunson et al. [2008], Dunson [2009], Petrone et al. [2009], which often do not share this property. This is particularly important for applications to mixture models because otherwise computations can be quite intensive.

3.4.6 Comparison with different approaches

In recent literature, there have been many proposals of generalizations of the Dirichlet process, particularly, dependent Dirichlet processes. Sev-

eral such proposals are discussed in Chapter 2. These approaches exploit marginal conditional independence. One considers a collection of random variables $\{Y_x, x \in \mathcal{X}\}$ and assumes that they are conditionally independent, that is, for any $x_1, \ldots, x_m \in \mathcal{X}$, ones assumes $Y_{x_1}, \ldots, Y_{x_m} \mid P_{x_1}, \ldots, P_{x_m} \sim \prod_{i=1}^{m} P_{x_i}(\cdot)$. Then, a prior is given on the family of random distributions $\{\mathbf{P}_x, x \in \mathcal{X}\}$, such that the $\mathbf{P}_x$'s are dependent.

However, in such approaches, $\{\mathbf{P}_x, x \in \mathcal{X}\}$ is not necessarily a random conditional, since x may not be random. In particular, since the covariate may be non random, no σ-algebra on $\mathcal{X}$ is considered, and thus, measurability with respect to $\mathcal{B}_X$ a.s. is not required. If measurability with respect $\mathcal{B}_X$ a.s. is satisfied, this is a model on the random conditionals and does not induce a prior on the random joint distribution of (X, Y).

Instead, our approach gives a prior on the marginal-conditional pair and induces a prior on the joint. For a Dirichlet process with non atomic base measure, the random conditionals are independent and degenerate a.s. We are extending this by allowing non degenerate conditionals, but we will assume independence. A further extension would allow dependence among the random conditionals through a dependent Dirichlet process MacEachern [1999] if measurability with respect to $\mathcal{B}_X$ a.s. is satisfied. However, some properties will be lost. For example, for a DDP, we would lose conjugacy, and the model would become much more complex, and using the Hierarchical DP Teh et al. [2006] or the Nested DP Rodriguez and Dunson [2011] would remove dependence on x in the base measures for the conditionals.

Notice that the distribution of the conditional also as a random function of X is $\mathbf{P}_{Y|X}(\cdot|X) \sim \sum_{i=1}^{\infty} w_i \delta_{\mathbf{P}_{Y|X}(\cdot|\tilde{X}_i)}$. This resembles the prior for the Nested Dirichlet process, but is not directly comparable since $\mathbf{P}_{Y|X}(\cdot|X)$ is a different object than $\{\mathbf{P}_x, x \in \mathcal{X}\}$.

3.5 Example

We provide an illustration of the properties of the EDP prior in an application to mixture models. The problem we consider is comparing dif-

ferent schools based on national test scores. The dataset we analyse contains two different test scores for students in 65 inner-London schools. The first score is based on the London Reading Test (LRT), taken at age 11, and the second is a score derived from the Graduate Certificate of Secondary Education (GCSE) exams in a number of different subjects, taken at age 16. Taking into account earlier LRT scores can give a sense of the "value added" for each school. To answer the question of which schools are most effective, we consider modeling the relationship between LRT and GCSE for all schools. The data are available at `http:// www.stata-press.com/data/mlmus.html`. School number 48 is dropped from the dataset since only 2 students were observed.

Rabe-Hesketh and Skrondal [2005] (Chapter 4) study the following multilevel parametric model where Y_{ij} and X_{ij} represent, respectively, the GCSE and LRT score for student i in school j:

$$\begin{aligned} Y_{ij} \mid \beta_{0j}, \beta_{1j}, x_{ij} &\overset{ind}{\sim} N(\beta_{0j} + \beta_{1j} x_{ij}, \sigma^2), \\ \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} &\overset{iid}{\sim} N_2\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \Sigma_\beta \right), \end{aligned} \tag{3.18}$$

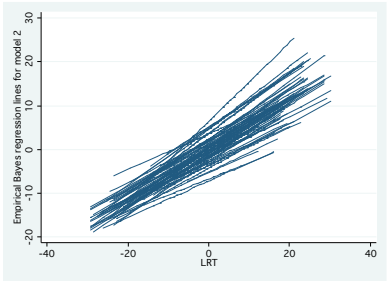
where β_{0j} and β_{1j} are independent of X_{ij}. The interest is in estimating the school specific coefficients $\beta_j = (\beta_{0j}, \beta_{1j})$. The intercept is interpreted as the school mean of GCSE scores for the students with the average LRT score of 0. The competitiveness of the school is captured by the school specific slope. Schools with greater slopes are competitive; more "value" is added for students with higher LRT scores. Schools with a slope of 0 are non-competitive; the performance of students is homogeneous regardless of how the students scored on the LRT. If parents are to choose the best school for their children, both average "value added" and competitiveness are important.

Maximum likelihood estimates of the parameters of the mixing distribution (Rabe-Hesketh and Skrondal [2005]) give $\hat{\beta}_0 = -.115$, with standard error $SE(\hat{\beta}_0) = .0199$, and $\hat{\beta}_1 = .55$, with $SE(\hat{\beta}_1) = .3978$, and

estimated covariance matrix:

$$\hat{\Sigma}_\beta = \begin{bmatrix} 9.04 & .18 \\ .18 & .0145 \end{bmatrix}.$$

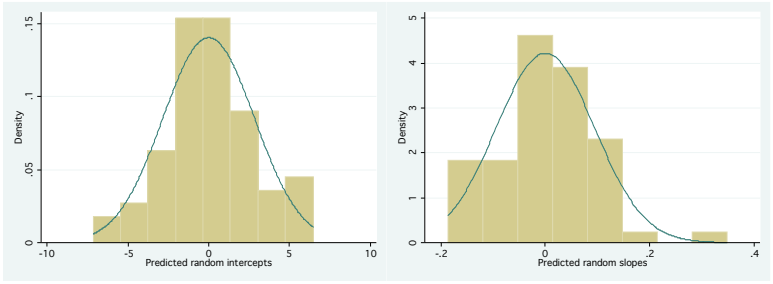
Empirical Bayes predictions of school specific intercept and slope were then obtained; figures (3.1a) and (3.1b) show the plots of estimated regression lines for each school and ranking of schools based on the intercept.



(a) Empirical Bayes Predictions of school-specific regression line

(b) Ranking of Schools

Figure 3.1: Results of Linear Mixed Effects model



(a) Predicted random intercept

(b) Predicted random slope

Figure 3.2: Assessing the model

By visual inspection of the histograms of the empirical Bayes estimates in figures (3.2a) and (3.2b), for the intercept and especially the slope, a

normal distribution does not fit well. This may be due to the fact that there are only 65 schools, that the normality assumption does not hold or a combination of the two. To enlarge the class of models, we can consider modelling the mixing distribution of the intercept and slope nonparametrically. A pitfall of model (3.18) is that it assumes the same variability for all schools. In fact, the wide range of the naive OLS estimates of within school variance (not shown) supports a model which allows for school-specific variance.

Bayesian nonparametric extensions of this model would assign a DP prior on the mixing distribution of the (β_{0j}, β_{1j})'s (a DP-location mixture), assuming the same variance σ^2 for each school, or model school specific variances σ_j^2, with a DP prior for the latent distribution of $(\beta_{0j}, \beta_{1j}, \sigma_j^2)$ (DP scale-location mixture). The EDP is an intermediate choice. It may model clusters of schools that share the same variance, with different β's inside each cluster. We assume that

$$\begin{aligned} Y_{ij}|x_{ij}, \beta_j, \sigma_j^2 &\overset{ind}{\sim} N(\beta_{0j} + \beta_{1j} x_{ij}, \sigma_j^2), \\ (\beta_j, \sigma_j^2)|P_{\beta,\sigma^2} &\overset{iid}{\sim} P_{\beta,\sigma^2}, \\ \mathbf{P}_{\beta,\sigma^2} &\sim \text{EDP}(\alpha, \mu), \end{aligned}$$

where $\beta_j = (\beta_{0j}, \beta_{1j})$ and the parameters of the EDP are specified as $\alpha = \alpha_{\sigma^2} P_{0,\sigma^2}$ and $\mu(\cdot, \sigma^2) = \mu_\beta(\sigma^2) P_{0,\beta|\sigma^2}(\cdot|\sigma^2)$ for all $\sigma^2 \in \mathbb{R}_+$.

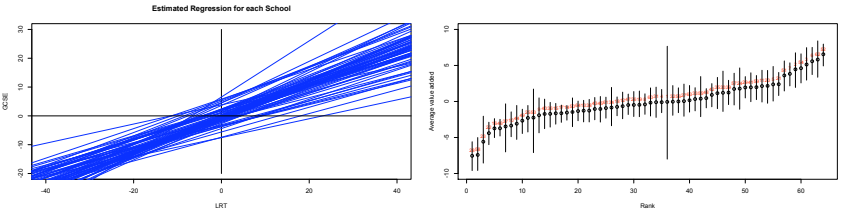
In the analysis reported below, we fixed the baseline measures $P_{0\sigma}$ as an Inverse-Gamma, with rate and shape parameters, respectively, 8 and 385, and $P_{0,\beta|\sigma^2}(\cdot|\sigma^2)$ as a bivariate Normal, $N_2(\mu_0, c_0\, \sigma^2\, \Sigma_0)$, with $\mu_0 = [0, .5]'$, $c_0 = 1/20$ and

$$\Sigma_0 = \begin{bmatrix} 9 & 3/16 \\ 3/16 & 1/64 \end{bmatrix}.$$

Notice that if the precision parameter $\alpha_{\sigma^2} \approx 0$, we get back to a DP location mixture, and if the precision parameters $\mu_\beta(\sigma^2) \approx 0$ for all $\sigma^2 \in \mathbb{R}_+$, we get a DP scale-location mixture. Thus, with an EDP prior we can express uncertainty between homoskedasticity and heteroskedasticity.

We model uncertainty about α_{σ^2} and $\mu_\beta(\sigma^2)$ through Gamma hyper-priors:
$\alpha_{\sigma^2} \sim \text{Gamma}(u_\alpha, v_\alpha)$, where we choose $u_\alpha = 2$ and $v_\alpha = 1$, and for all $\sigma^2 \in \mathbb{R}_+$ $\mu_\beta(\sigma^2) \overset{iid}{\sim} \text{Gamma}(u_{\mu_\beta}, v_{\mu_\beta})$, with $u_{\mu_\beta} = 2$ and $v_{\mu_\beta} = 1$.

The MCMC scheme to compute posterior distributions is based on the algorithm 6 described in Neal [2000], which is a Metropolis-Hastings algorithm with candidates drawn from the prior. Resampling the precision parameters is done by introducing a latent beta-distributed variable, as described in Escobar and West [1995]. The number of iterations is set up to $20,000$ with 10% of burn-in. Looking at the trace and autocorrelation plots, convergence appears reached for the β's in all schools and for σ^2's in most schools. The results are summarized in Figures (3.3a) and (3.3b), which display the estimated regression line for each school and the ranking of schools based on average "value added " with empirical quantiles.



(a) Estimated regression line for each school (b) Ranking of Schools based on average value added with empirical quantile

Figure 3.3: Results of EDP model

The MCMC posterior expectation of α_{σ^2} is 2.5, and Figure (3.4) depicts the estimated posterior values of $\mu_\beta(\sigma^2)$ for different values of σ^2.

Neither $\alpha_{\sigma^2} \approx 0$ nor $\mu_\beta(\sigma^2) \approx 0$ for all σ^2, and interestingly, the estimated values of $\mu_\beta(\sigma^2)$ are high for values of σ^2 which are more likely a posteriori, and close to zero for unlikely values of σ^2. Thus, the results favor a model which allows for homoskedasticity among some schools with a more likely value σ^2 and some outlying schools with abnormally large or

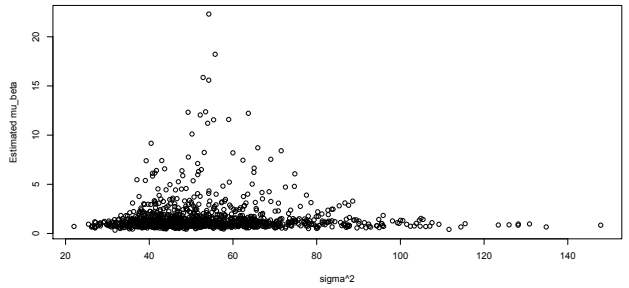


Figure 3.4: Estimated posterior values of $\mu_\beta(\sigma^2)$ for different values of σ^2

small variances.

3.6 Discussion

We have proposed an *enrichment* of the DP starting from the idea of enriched conjugate priors. The advantages of this process are that it allows for more flexible specification of prior information, includes the DP as a special case, and retains some desirable properties including conjugacy and the fact that it can be constructed from an enriched urn scheme. The disadvantages include the difficulty in obtaining a closed form for the distribution of the joint probability over a given set and for the distribution of the marginal probability over a measurable subset of $\mathcal{Y}$. Using an EDP as the prior for the distribution of a random vector, Z, implies one has to determine a partition of Z into two groups and an ordering defining which group comes first. The "two-level" clustering resulting from the EDP introduces a clear asymmetry based on the partition and ordering chosen, and how to choose them depends on the application. There may be a natural ordering or partition and/or computational reasons, including decomposition of the base measure, for choosing the partition and ordering. In our example, we partitioned the random vector $(\beta_0, \beta_1, \sigma^2)$ into the two groups, (σ^2) and (β_0, β_1), with σ^2 chosen first due to uncertainty

in homoskedasticity and decomposition of the conjugate normal-inverse gamma base measure. One may also examine all plausible and interesting partitions and orderings.

We have focused on the partition of the random vector into two groups, but most results could be extended to any finite partition of the random vector, although this would of course imply a further nested structure. In the next chapter, we examine the implied clustering structure in regression settings when the joint model is an EDP mixture. Other extensions could include exploring if other conjugate nonparametric priors whose finite dimensionals are standard conjugate priors can be generalized starting from enriched conjugate priors, such as extension of the enriched distribution, mentioned in the Remark 2, to an enriched bivariate Neutral to the Right Processes.

We hope that having explored these features can shed light on potentialities and limitations and encourage further developments in constructing more flexible priors for a random probability measure on $\mathbb{R}^p$.

Chapter 4

Enriched Dirichlet process mixtures for regression

Flexible covariate-dependent density estimation can be achieved by modelling the joint density of the response and covariate as a Dirichlet process mixture. An appealing aspect of this approach is that computations are relatively easy. In this chapter, we examine the predictive performance of these models with an increasing number of covariates. Even for a moderate number of covariates, we find that the likelihood for x tends to dominate the posterior of the latent random partition, degrading the predictive performance of the model. To overcome this, we propose to replace the Dirichlet process with the Enriched Dirichlet process. Our proposal maintains a simple allocation rule, so that computations remain relatively simple. Advantages are shown through both predictive equations and examples, including an application to diagnosis Alzheimer's disease.

This chapter contains joint work with Sonia Petrone and will be submitted for publication shortly. We would like to thank David B. Dunson for bringing the problem to our attention.

4.1 Introduction

Dirichlet process mixture models are important tools for density estimation. Theoretical properties such as strong and weak consistency are satisfied for a large class of data-generating densities (Ghosal et al. [1999], Ghosal and van der Vaart [2001], Ghosal and van der Vaart [2007],Tokdar [2006], Walker et al. [2007], Wu and Ghosal [2008], Wu and Ghosal [2010], Tokdar [2011]), and efficient computational procedures are well known (MacEachern [1994], Ishwaran and James [2001], Neal [2000], Papaspiliopoulos and Roberts [2008], Kalli et al. [2011]). From an interpretative perspective, a further appealing aspect is the clustering implied by the DP.

DP mixture models can be extended to treat the problems of estimating a regression function and a conditional density by simply augmenting the observations to include the response and covariates (y, x) and modeling the joint density through a DP mixture. The regression function and conditional density estimates are obtained from the estimate of the joint density, an idea which is similarly employed in classical kernel regression methods (Scott [1992], Chapter 8).

The joint approach based on DP mixtures was first introduced by Müller et al. [1996], and subsequently studied by many others including Kang and Ghosal [2009], Shahbaba and Neal [2009], Hannah et al. [2011], Park and Dunson [2010], and Müller and Quintana [2010]. The implied latent clustering of the DP is particularly useful in the regression setting. In particular, if the kernel for y given x is the standard linear regression model, the DP model uses simple linear regression models as building blocks and partitions the observed subjects into clusters, where within cluster, the linear regression model provides a good fit. Even though within cluster, the model is parametric, globally, a wide range of complex distributions can describe the joint distribution, leading to a flexible model for both the regression function and the conditional distribution.

Recent literature contains many generalizations of the DP to define a flexible model for covariate-dependent density estimation based on a

conditional approach. In such models, the conditional density of $Y|x$ is modelled directly, where $f(y|x,\theta)$ is parametric and the parameter θ conditional to x has an unknown distribution, $\mathbf{P}_x$, depending on x. A prior is then given on the family of distributions $\{\mathbf{P}_x, x \in \mathcal{X}\}$ such that the $\mathbf{P}_x$'s are dependent. Examples for the law of $\mathbf{P}_\mathcal{X}$, which include MacEachern [1999], MacEachern [2000], Griffin and Steele [2006], Dunson and Park [2008], Ren et al. [2011], Chung and Dunson [2009], and Rodriguez and Dunson [2011], are given in Section 2.3 of Chapter 2.

Such models based on a conditional approach can approximate a wide range of response distributions that may change flexibly with the covariate. However, computations are often quite burdensome. One of the reasons the model examined here is so powerful is its simplicity. Together, the joint approach and the clustering of the DP provide a built-in technique to allow for changes in the response distribution across the covariate space, yet it is simple and generally less computationally intensive than the nonparametric conditional models based on dependent DPs.

Other regression techniques focus on flexibly modelling the regression function, but do not provide a flexible model for the conditional distribution. Many of these techniques, such as splines or multivariate extensions of splines, rely on partitioning the covariate space into groups. These techniques suffer heavily from the curse of dimensionality, requiring an increasingly higher number of subregions of the covariate space as p, the dimension of X, increases, fueling the need for larger sample sizes to obtain reliable estimates (Kang and Ghosal [2009]). Instead, the joint DP mixture model is able to avoid this problem by partitioning the observed subjects into groups instead of the covariate space. Unfortunately, other, more subtle issues arise with increasing p.

This random allocation of subjects into groups is driven by the need to obtain a good approximation of the joint distribution of Y and X. This means partitions of subjects with similar covariates, as measured by the likelihood for x, and similar relationship between the response and covariate, as measured by the likelihood for $y|x$, will have higher posterior mass. However, as p increases, the likelihood for x tends to dominate

the posterior of the random partition, so that clusters are based solely on similarity in the covariate space. This problem was first brought to our attention by Professor David B. Dunson through personal communication, and discussed, but not fully developed, in an unpublished manuscript by Dunson et al. [2011].

In many applications, the density of X may be complex and require several kernels for a good approximation, while the density of Y given x may be more stable. This is particularly common in high-dimensions, when often, for statistical and computational reasons, simple kernels, assuming independence of the covariates, are used. If the covariates are dependent, many kernels will be needed to approximate the dependency in the density of X. Generally, a larger p results in a higher degree of multicollinearity. Thus, if there are clusters of subjects with a similar behavior of y given x, but the covariates exhibit multicollinearity within cluster, the partition will consist of many sub-clusters due to the dominance of likelihood for x. This may cause less reliable estimates and large credible intervals due to small sample sizes within cluster. To address this issue, one may want to allow for more x-clusters.

In other applications, this behavior of the partition structure may be unappealing when the response of subjects belonging to the same cluster in the covariate space may exhibit multiple types of behavior or other departures from the local model for $Y|x$. In this case, subjects may belong to the same x-cluster but possibly different y-clusters to obtain a good approximation to the conditional density of $y|x$. When p is small, these subjects will be placed in different clusters, but, when p is large, these subjects will be forced to belong to the same cluster. This may result in poor and inaccurate predictive density estimates and credible intervals. To bypass this problem, one may want to allow further y-clusters.

These problems suggest that for moderate to large p, a different clustering structure for the marginal of X and the regression of Y on x may be desirable to allow of the impact of x on Y to influence the clustering structure, improving predictive estimates. In this chapter, we propose to replace the DP with the Enriched Dirichlet process (EDP) developed

in Chapter 3, allowing a nested clustering structure that can overcome these issues. An alternative proposal is discussed in Petrone and Trippa [2009] and Dunson et al. [2011], where they suggest the use of a partially hierarchical Dirichlet process. In a Bayesian nonparametric framework, several extensions of the Dirichlet process have been proposed to allow local clustering (Dunson et al. [2008], Dunson [2009], Petrone et al. [2009]). However, the greater flexibility is often achieved at price of more complex computations. Instead, our proposal maintains a simple, analytically computable, allocation rule, and therefore, computations are a straightforward extension of those used for the joint DP mixture model.

This chapter is organized as follows. In Section 4.2, we review the joint DP mixture model, its covariate-dependent random partition model, and carefully examine the predictive performance. We discuss two situations where prediction could be improved and for the remainder of the chapter, focus on one, when the density of X requires many kernels for a good approximation. In Section 4.3, we propose a joint EDP mixture model, discuss its covariate-dependent random partition model, and emphasize the predictive improvements for the problem of interest. Section 4.4 covers computational procedures. We provide a simulated example in Section 4.5 to demonstrate how the EDP model can lead to more efficient estimators by making better use of information contained in the sample. Finally, in Section 4.6, we apply the model to predict Alzheimer's Disease status based on measurements of various brain structures.

4.2 Joint DP mixture model

Müller, Erkanli, and West [1996] were the first to propose modelling the joint distribution of (X, Y) with a DP mixture model in order to obtain inference on the distribution of $Y|X = x$. They assume the distribution of (X, Y) is a DP mixture of multivariate normals and use a conjugate Normal Inverse Wishart prior for the base measure of the DP.

Shahbaba and Neal [2009] extend this model by re-parametrizing in terms of the parameters of the marginal of X and the conditional of $Y|x$.

This re-parametrization allows for two important extensions. First, the distribution of $Y|x$ can now have any parametric form, and thus, the model can handle other response types such as discrete Y. Secondly, the method can now handle high-dimensional covariates.

Indeed, in the parametrization of Müller, Erkanli, and West, the slopes of the local regression lines are determined by the local covariance matrices. Therefore, if the dimension of X is p, to have a flexible model for the local regression lines, we need to assign a prior for the full $p+1$ by $p+1$ covariance matrix, which poses both computational and statistical difficulties. The computational cost of computing and sampling from the posterior greatly increases with large p; in particular, there are $(p+1)(p+2)/2$ parameters for the $p+1$ by $p+1$ covariate matrix. Also, assigning a flexible prior that incorporates prior information for the full covariance matrix can be statistically difficult due to the positive semi-definite requirement.

Shahbaba and Neal assume independence among the covariates locally, i.e. the covariance matrix of the kernel on $\mathcal{X}$ is diagonal. Thus, a prior for the covariance matrix, now reduces to a prior for the p variances of the covariates, which greatly eases both the computational and statistical issues. Furthermore, the model for the local linear regression is still flexible. Note that even though, within each component, we assume independence of the covariates, globally, there is dependence. Local independence of the covariates also allows for easy inclusion of discrete or other types of covariates.

Shahbaba and Neal focus on the case when Y is categorical and the local model for $Y|x$ is a multinomial logit. Hannah et al. [2011] extend this approach by assuming that, locally, the conditional distribution of $Y|x$ belongs to the class of generalized linear models (GLM), that is, the distribution of the response belongs to the exponential family and the mean of the response can be expressed a function of a linear combination of the covariates. An interesting contribution is their study of asymptotic properties of the model. As Shahbaba and Neal, they also consider local independence of the covariates.

Kang and Ghosal [2009] study the model using an empirical Bayes ap-

proach approach for inference and through simulated examples, compare their results with standard regression techniques, such as splines and multivariate extensions of splines. They find that when the model assumptions hold, their approach leads to significantly smaller estimation error, with a pronounced effect in higher-dimensions.

The model, in full generality, can be described as follows:

$$\begin{aligned} Y_i|x_i,\theta_i &\overset{ind}{\sim} F_y(\cdot|x_i,\theta_i), \\ X_i|\psi_i &\overset{ind}{\sim} F_x(\cdot|\psi_i), \\ (\theta_i,\psi_i)|P &\overset{iid}{\sim} P, \\ \mathbf{P} &\sim \mathrm{DP}(\alpha P_{0Y}\times P_{0X}). \end{aligned} \tag{4.1}$$

Integrating out the subject-specific parameters, θ_i, ψ_i, the model for the joint density is

$$f_P(y_i,x_i)=\sum_{j=1}^{\infty} w_j K(y_i;x_i,\tilde{\theta}_j)K(x_i;\tilde{\psi}_j),$$

where

$$\mathbf{P}=\sum_{j=1}^{\infty} w_j \delta_{(\tilde{\theta}_j,\tilde{\psi}_j)},$$

and the kernels $K(y;x,\theta)$ and $K(x;\psi)$ are the densities associated to $F_y(\cdot|x,\theta)$ and $F_x(\cdot|\psi)$.

4.2.1 Random partition

One of the crucial features of this model is the dimension reduction and clustering obtained due to the almost sure discreteness of $\mathbf{P}$. In fact, it is often convenient to reparametrize in terms of the random partition of subjects into clusters and the unique values of the subject-specific parameters. The notation for the random partition is consistent with that used in Chapter 2. In particular, the partition of the n subjects is represented by $\rho_n = (s_1,\ldots,s_n)$, with $s_i = j$ if the parameter of subject i is j^{th} unique value observed. The unique values of subject-specific parameters

is denoted by $(\theta^*, \psi^*) = (\theta_j^*, \psi_j^*)_{j=1}^k$, where k is the number of unique values. The number of subjects with the j^{th} unique value is denoted n_j, and $S_j = \{i : s_i = j\}$ is the set of subject indices in the j^{th} cluster. Furthermore, we use the notation $y_j^* = \{y_i\}_{i \in S_j}$ and $x_j^* = \{x_i\}_{i \in S_j}$.

By jointly modeling Y and X, we introduce dependency between x and ρ_n. Park and Dunson [2010] examine the distribution of the covariate-dependent random partition. In particular, given the covariates and the unique parameters,

$$p(\rho_n | x_{1:n}, \psi^*) \propto \alpha^k \prod_{j=1}^{k} \Gamma(n_j) \prod_{i \in S_j} K(x_i; \psi_j^*). \tag{4.2}$$

Equation (4.2) shows that given $x_{1:n}$ and ψ^*, partitions containing clusters of subjects with covariates that are well described by $K(\cdot|\psi_j^*)$ are encouraged. When P_{0X} is the conjugate prior, the x-parameters, (ψ_j^*), can be analytically integrated out, since they are often not of interest in the analysis. Following this approach, the covariate random partition model is obtained by integrating the likelihood of x_j^* with respect to P_{0X}:

$$p(\rho_n | x_{1:n}) \propto \alpha^k \prod_{j=1}^{k} \Gamma(n_j) g_x(x_j^*),$$

where g_x is the marginal likelihood of x_j^* under the base measure:

$$g_x(x_j^*) = \int_{\Psi} \prod_{i \in S_j} K(x_i; \psi) dP_{0X}(\psi).$$

Independently, Müller and Quintana [2010] construct a similar covariate-dependent random partition model, but were motivated by directly modifying the cohesion term of a product partition model by a factor that encourages clusters with similar covariates.

The posterior of the covariate random partition, given also θ^* and ψ^*, is

$$p(\rho_n | x_{1:n}, y_{1:n}, \theta^*, \psi^*) \propto \alpha^k \prod_{j=1}^{k} \Gamma(n_j) \prod_{i \in S_j} K(x_i; \psi_j^*) K(y_i; x_i, \theta_j^*). \tag{4.3}$$

Therefore, integrating out the unique parameters, the posterior of the covariate random partition model is

$$p(\rho_n|x_{1:n}, y_{1:n}) \propto \alpha^k \prod_{j=1}^{k} \Gamma(n_j) g_x(x_j^*) g_y(y_j^*|x_j^*), \tag{4.4}$$

where g_y is defined, similar to g_x, as

$$g_y(y_j^*|x_j^*) = \int_\Theta \prod_{i \in S_j} K(y_i; x_i, \theta) dP_{0Y}(\theta).$$

From (4.3) and (4.4), we see that given the data, subjects are clustered in groups with similar behaviour in the covariate space and similar relationship with the response. However, even for moderate p the likelihood for x tends to dominate the posterior of the random partition, so that clusters are determined only by similarity in the covariate space. This is particularly evident when the covariates are assume to be independent locally, i.e.

$$K(x_i; \psi_j^*) = \prod_{h=1}^{p} K(x_{i,h}; \psi_{j,h}^*).$$

Clearly, for large p, the scale and magnitude of changes in $\prod_{h=1}^{p} K(x_{i,h}; \psi_{j,h}^*)$ will wash out any information given in the univariate likelihood $K(y_i; \theta_j^*, x_i)$. This behavior is particularly undesirable if the data of interest falls into one of the two cases.

The first case consists of datasets where the distribution of X displays many departures from $F_x(\cdot; \psi)$. This behavior is common in high-dimensions due to the fact that for reasons previously mentioned, the covariates are assumed independent locally, yet as p increases, the degree of multicollinearity typically also increases. Many departures from $K(x; \psi)$ will cause the number of components to grow, yet the conditional distribution of Y may be more stable and require much less components.

For a simple example demonstrating how the number of components needed to approximate marginal of X can blow up with p, imagine X is uniformly distributed on a cuboid of side length $r > 1$. Consider approxi-

mating

$$f_0(x) = \frac{1}{r^p}\mathbf{1}(x \in [0,r]^p)$$

by

$$f_k(x) = \sum_{j=1}^{k} w_j N_p(x; \mu_j, \sigma_j^2 I_p).$$

Since the true distribution of x is uniform on the cube $[0,1]^p$, to obtain a good approximation, the weighted components must place most of their mass on values of x contained in the cuboid. Let $B_\sigma(\mu)$ denote a ball of radius σ centered at μ. If a random vector V is normally distributed with mean μ and variance $\sigma^2 I_p$, then for $0 < \epsilon < 1$,

$$P(V \in B_{\sigma z(\epsilon)}(\mu)) = 1 - \epsilon,$$

where

$$z(\epsilon)^2 = (\chi_p^2)^{-1}(1-\epsilon),$$

i.e. the square of $z(\epsilon)$ is the $(1-\epsilon)$ quantile of the chi-squared distribution with p degrees of freedom. For small ϵ, this means that the density of V places most of its mass on values contained in a ball of radius $\sigma z(\epsilon)$ centered at μ. For $\epsilon > 0$, define

$$\tilde{f}_k(x) = \sum_{j=1}^{k} w_j N(x; \mu_j, \sigma_j^2 I_p) * \mathbf{1}(x \in B_{\sigma_j z(\epsilon_j)}(\mu_j)),$$

where $\epsilon_j = \epsilon/(kw_j)$. Then, $\tilde{f}_k$ is close to f_k (in the L_1 sense):

$$\begin{aligned}\int_{\mathbb{R}^p} |f_k(x) - \tilde{f}_k(x)|dx &= \int_{\mathbb{R}^p} \sum_{j=1}^{k} w_j N(x; \mu_j, \sigma_j^2 I_p) * \mathbf{1}(x \in B^c_{\sigma_j z(\epsilon_j)}(\mu_j))dx, \\ &= \sum_{j=1}^{k} w_j \frac{\epsilon}{kw_j} = \epsilon.\end{aligned}$$

And, for $\tilde{f}_k$ to be close to f_0, the parameters μ_j, σ_j, w_j need to be chosen so that the balls $B_{\sigma_j z(\epsilon/(kw_j))}(\mu_j)$ are contained in the cuboid. That means that centers of the balls are contained in the cuboid,

$$\mu_j \in [0,r]^p, \tag{4.5}$$

with further constraints on σ_j^2 and w_j, so that the radius is small enough. In particular,

$$\sigma_j z \left(\frac{\epsilon}{k w_j} \right) \leq \min(\mu_1, r - \mu_1, \ldots, \mu_p, r - \mu_p) \leq \frac{r}{2}. \tag{4.6}$$

However, as p increases the volume of the cuboid goes to infinity, but the volume of any ball $B_{\sigma_j z(\epsilon/(kw_j))}(\mu_j)$ defined by (4.5) and (4.6) goes to 0 (see Clarke et al. [2009], Section 1.1). Thus, just to reasonably cover the cuboid with the balls of interest, the number of components will increase dramatically, and more so, when we consider the approximation error of the density estimate. Now, as an extreme example, imagine that $f_0(y|x)$ is a linear regression model. Even though one component is sufficient for $f_0(y|x)$, a large number of components will be required to approximate $f_0(x)$, particularly as p increases.

The second case where dominance of x in partition structure may be problematic consists of datasets where the response of subjects belonging to the same cluster in the covariate space may exhibit multiple types of behavior or display other departures from the local model $K(y; x, \theta)$. In order to obtain a good approximation of the response distribution, the x-clusters would need to be divided into sub-clusters. However, this may not occur if p is large due to dominance of x in determining the clustering structure.

4.2.2 Posterior of the unique parameters

Next, we examine how the dominance of x in the partition structure effects the posterior of the unique parameters, which, in turn, has important implications for the prediction. *Aposteriori* the cluster parameters, (θ_j^*, ψ_j^*), are independent,

$$p(\theta^*, \psi^* | y_{1:n}, x_{1:n}, \rho_n) = \prod_{j=1}^{k} p(\theta_j^* | y_j^*, x_j^*) p(\psi_j^* | x_j^*),$$

with posterior density

$$p(\theta_j^*|y_j^*, x_j^*) \propto p_{0Y}(\theta_j^*) \prod_{i \in S_j} K(y_i; x_i, \theta_j^*),$$
$$p(\psi_j^*|x_j^*) \propto p_{0X}(\psi_j^*) \prod_{i \in S_j} K(x_i; \psi_j^*),$$

where, p_{0Y} and p_{0X} are the densities of P_{0Y} and P_{0X}. If P_{0Y} and P_{0X} are the conjugate priors, then *aposteriori* the prior parameters of (θ_j^*, ψ_j^*) are updated based on subjects in S_j.

In the first situation, the model may require many kernels to approximate the density of x with a small number of individuals within each cluster. In this case, the posterior for θ_j^* will be based on small sample sizes, leading to a flat posterior with an unreliable posterior mean and large influence of the prior.

In the second, cluster j may contain subjects whose density cannot be described by $K(y; x, \theta_j^*)$, but they are forced to be in the same cluster because of similarity of their covariates. In this case, posterior inference of θ_j^* will be poor due to inaccurate modelling.

4.2.3 Covariate-dependent urn scheme

Our aim is prediction of the mean and conditional density of the response for a new subject. Given ρ_n and (θ^*, ψ^*), the prediction and predictive density at a new value of x can be computed analytically. This computation relies on the predictive distribution of s_{n+1}, which, also given (θ^*, ψ^*), is

$$s_{n+1}|\rho_n, \psi^*, x_{1:n+1} \sim \frac{w_{k+1}^*(x_{n+1})}{c_0}\delta_{k+1} + \sum_{j=1}^{k} \frac{w_j^*(x_{n+1})}{c_0}\delta_j, \tag{4.7}$$

where $c_0 = p(x_{n+1}|\rho_n, \psi^*) * (\alpha + n)$ is a normalizing constant,

$$w_j^*(x_{n+1}) = n_j K(x_{n+1}; \psi_j^*) \text{ for } j = 1, \ldots, k,$$

and

$$w_{k+1}^*(x_{n+1}) = \alpha g_x(x_{n+1}).$$

Again, the parameters ψ^* may be analytically integrated out if P_{0X} is conjugate. In particular, $K(x_{n+1}; \psi_j^*)$ is integrated with respect to the posterior of ψ_j^* given x_j^*, resulting in a covariate-dependent urn scheme similar to (4.7) with weights for $j = 1, \ldots, k$ defined by

$$w_j^{*\prime}(x_{n+1}) = n_j g_x(x_{n+1}|x_j^*),$$

and a normalizing constant of $c_0' = p(x_{n+1}|\rho_n, x_{1:n}) * (\alpha + n)$, where

$$g_x(x_{n+1}|x_j^*) = \int_\Psi K(x_{n+1}; \psi) dP(\psi|x_j^*).$$

Note that this urn scheme is a generalization of the classic Pólya urn scheme that allows the probabilities of cluster membership to depend on the covariate, where the new subject is placed cluster j if his covariate is similar to the covariates of subjects in cluster j as measured by the predictive density $g_x(\cdot|x_j^*)$. See Park and Dunson [2010] for more details.

4.2.4 Prediction

We now have all tools needed to compute the predictive estimates. Under the squared error loss function, the prediction of y_{n+1} for a new subject with a covariate of x_{n+1} is

$$\begin{aligned} \mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}] &= \sum_{\mathcal{P}_n} \int_{\Theta^k} \int_{\Psi^k} [\ldots] dP(\rho_n, \theta^*, \psi^*|y_{1:n}, x_{1:n}), \\ [\ldots] &= \frac{w_{k+1}^*(x_{n+1})}{c_1} E_{G_y}[Y_{n+1}|x_{n+1}] + \sum_{j=1}^{k} \frac{w_j^*(x_{n+1})}{c_1} E_{F_y}[Y_{n+1}|x_{n+1}, \theta_j^*], \end{aligned} \tag{4.8}$$

where $c_1 = p(x_{n+1}|x_{1:n}) * (\alpha + n)$, $\mathcal{P}_n$ denotes the set of partitions of the first n integers, and

$$G_y(\cdot|x) = \int_\Theta F_y(\cdot|x, \theta) dP_{0Y}(\theta).$$

Similarly, the predictive density at y for a new subject with a covariate

of x_{n+1} is

$$f(y|y_{1:n}, x_{1:n+1}) = \sum_{\mathcal{P}_n} \int_{\Theta^k} \int_{\Psi^k} [\ldots] dP(\rho_n, \theta^*, \psi^*|y_{1:n}, x_{1:n}),$$

$$[\ldots] = \frac{w^*_{k+1}(x_{n+1})}{c_1} g_y(y|x_{n+1}) + \sum_{j=1}^{k} \frac{w^*_j(x_{n+1})}{c_1} K(y_{n+1}; x_{n+1}, \theta^*_j). \quad (4.9)$$

For example, when $K(y; x, \theta) = N(y; \underline{X}\beta, \sigma^2)$ and the prior for (β, σ^2) is the multivariate normal-inverse gamma with parameters (β_0, C, a_y, b_y), (4.8) is

$$\frac{w^*_{k+1}(x_{n+1})}{c_1} \underline{X}_{n+1}\beta_0 + \sum_{j=1}^{k} \frac{w^*_j(x_{n+1})}{c_1} \underline{X}_{n+1}\beta^*_j, \quad (4.10)$$

and (4.9) is

$$\frac{w^*_{k+1}(x_{n+1})}{c_1} \mathcal{T}(y; \underline{X}_{n+1}\beta_0, W^{-1}_{n+1} \frac{b_y}{a_y}, 2a_y) + \sum_{j=1}^{k} \frac{w^*_j(x_{n+1})}{c_1} N(y; \underline{X}_{n+1}\beta^*_j, \sigma^{2*}_j), \quad (4.11)$$

where $\mathcal{T}(\cdot; \mu, \sigma^2, \nu)$ denotes the density of random variable, V, such that $(V - \mu)/\sigma$ has a t-distribution with ν degrees of freedom, and

$$W_{n+1} = 1 - \underline{X}_{n+1}(C + \underline{X}'_{n+1}\underline{X}_{n+1})^{-1}\underline{X}'_{n+1}.$$

Notice that given the partition and the unique parameters, the prediction or predictive density is a weighted average of the predictions within each cluster. By allowing for the urn scheme to depend on the covariate, the weights assigned to prediction within each cluster depend on the covariates. These covariate-dependent weights are important for prediction because cluster predictions associated with covariates similar to x_{n+1} will be given more weight in the overall prediction.

However, for moderate to large p, the posterior of the partition may favor clusters with similar x independent of $y|x$ behaviour, which can negatively effect both the prediction and predictive density. In the first situation, given the partition, the prediction will be an average over the

large number of within cluster predictions, which are based on small sample sizes. This will result in unreliable estimates with large prior influence and high variability. Furthermore, the measure which determines similarity of x_{n+1} and the j^{th} cluster will be too rigid. In the second situation, the prediction and the predictive density within cluster may not be flexible enough to capture the behaviour present in the data due to poor posterior inference of θ^* and incorrect modelling within cluster.

4.3 Joint EDP mixture model

In this section, we address the issues discussed in the previous section. We focus on the first problem, which considers datasets that require many kernels to approximate the density of X, a common issue in high-dimensions. The conditional density of $Y|x$, on the other hand, may be more stable. Thus, a local clustering of the subject-specific parameters $(\theta_i, \psi_i)_{i=1}^n$ is desirable. Recent proposals for local clustering (Dunson et al. [2008], Dunson [2009], Petrone et al. [2009]) could be used. However, computations are often quite burdensome. Instead, our proposal is to simply replace the DP with the more richly parametrized EDP, which is relatively easy from a computational perspective thanks to the analytically computable urn scheme of the EDP. The second problem discussed in the previous section can be addressed analogously by reversing the ordering of the (θ, ψ) in the definition of the EDP.

To clarify notation, we recall the definition of the EDP. The parameters consist of a finite measure α on Θ and a mapping $\mu(\cdot, \theta)$ such that for every $\theta \in \Theta$, it is a finite measure on Ψ and as a function of θ, it is α-integrable. In this chapter, the parameters will be reparametrized in terms of the base measure P_0 on $\Theta \times \Psi$, defined as

$$P_0(A \times B) = \int_A \frac{\mu(B, \theta)}{\mu(\Psi, \theta)} d\frac{\alpha(\theta)}{\alpha(\Theta)},$$

a precision parameter $\alpha_y = \alpha(\Theta)$ associated to θ and a collection of precision parameters $\alpha_x(\theta) = \mu(\Psi, \theta)$ for every $\theta \in \Theta$ associated to $\psi|\theta$. The EDP is defined by

1. $\mathbf{P}_Y \sim \mathrm{DP}(\alpha_y P_{0Y})$.
2. $\forall \theta \in \Theta$, $\mathbf{P}_{X|Y}(\cdot|\theta) \sim \mathrm{DP}(\alpha_x(\theta) P_{0X|Y}(\cdot|\theta))$.
3. $\mathbf{P}_{X|Y}(\cdot|\theta), \theta \in \Theta$ are independent among themselves.
4. $\mathbf{P}_Y$ is independent of $\left\{\mathbf{P}_{X|Y}(\cdot|\theta)\right\}_{\theta\in\Theta}$.

The law of the random joint $\mathbf{P}$ is obtained from the joint law of the marginal and conditionals through the mapping $(P_Y, P_{X|Y}) \to \int_{(\cdot)} P_{X|Y}(\cdot|\theta) dP_Y(\theta)$.

The proposed EDP mixture model for regression is

$$\begin{aligned} Y_i|x_i, \theta_i &\overset{ind}{\sim} F_y(\cdot|x_i, \theta_i), \\ X_i|\psi_i &\overset{ind}{\sim} F_x(\cdot|\psi_i), \\ (\theta_i, \psi_i)|P &\overset{iid}{\sim} P, \\ \mathbf{P} &\sim \mathrm{EDP}(\alpha, \mu). \end{aligned}$$

Integrating out $(\theta_1, \psi_1, \ldots, \theta_n, \psi_n)$, the model for the joint density is

$$f_P(x_i, y_i) = \sum_{j=1}^{\infty} \sum_{l=1}^{\infty} w_j w_{l|j} K(x_i; \tilde{\psi}_{l|j}) K(y_i; x_i, \tilde{\theta}_j),$$

where

$$\mathbf{P} = \sum_{j=1}^{\infty} \sum_{l=1}^{\infty} w_j w_{l|j} \delta_{(\tilde{\psi}_{l|j}, \tilde{\theta}_j)}.$$

4.3.1 Random partition

An important advantage of the EDP is the implied nested clustering. In particular, the EDP model partitions subjects in y-clusters and x-clusters within each y-cluster, allowing a more flexible local model for x within each y-cluster. An alternative proposal, which also induces a nested partition structure, is the partially hierarchical Dirichlet process (Petrone and Trippa [2009], Dunson et al. [2011]). This proposal, however, is more restrictive in the sense that there are only two precision parameters.

To describe the random partition model induced from the EDP, we need to introduce some notation. The partition can be described by the y-cluster memberships and x-cluster memberships, where $s_{y,i} = j$ if subject i is in the j^{th} y-cluster and $s_{x,i} = l$ if subject i is in the l^{th} x-cluster within its y-cluster. The cluster memberships are sorted in order of appearance, that is to say, the j^{th} y-cluster represents j^{th} y-species observed and the l^{th} x-cluster represents the l^{th} x-species observed among subjects in the same y-cluster. The set containing the indices of subjects in the j^{th} y-cluster will be represented by S_{j+}, and the set containing the indices of subjects in the l^{th} x-cluster within the j^{th} y-cluster will be represented by $S_{j,l}$. Let $\rho_n = (\rho_{n,y}, \rho_{n,x})$, $\rho_{n,y} = (s_{y,1}, ..., s_{y,n})$, $\rho_{n,x} = (s_{x,1}, ..., s_{x,n})$, and $\rho_{n_{j+},x} = (s_{x,i})_{i \in S_{j+}}$. The number of y clusters will be denoted by k with n_{j+} representing the number of subjects in j^{th} y-cluster, $j = 1, \ldots, k$, and the number of x-clusters in the j^{th} y-cluster will be denoted by k_j with $n_{j,l}$ representing the number of subjects in l^{th} x-cluster within j^{th} y-cluster, $l = 1, \ldots, k_j$ and $j = 1, \ldots, k$. The unique parameters will be denoted by $\theta^* = (\theta_j^*)_{j=1}^k$ and $\psi^* = (\psi_{1|j}^*, \ldots, \psi_{k_j|j}^*)_{j=1}^k$. Furthermore, we use the notation $y_j^* = \{y_i\}_{i \in S_{j+}}$, $x_j^* = \{x_i\}_{i \in S_{j+}}$ and $x_{j,l}^* = \{x_i\}_{i \in S_{j,l}}$.

Proposition 4.3.1 *The random partition model defined from the EDP is*

$$p(\rho_n) = \frac{\Gamma(\alpha_y)}{\Gamma(\alpha_y + n)} \alpha_y^k \prod_{j=1}^{k} \int_\Theta \alpha_x(\theta)^{k_j} \frac{\Gamma(\alpha_x(\theta))\Gamma(n_{j+})}{\Gamma(\alpha_x(\theta) + n_{j+})} dP_{0Y}(\theta) \prod_{l=1}^{k_j} \Gamma(n_{j,l}).$$

Proof. From independence of random conditional distributions among $\theta \in \theta$,

$$\begin{aligned} p(\rho_n, \theta^*) &= p(\rho_{n,y}) \prod_{j=1}^{k} p_{0Y}(\theta_j^*) p(\rho_{n,x} | \rho_{n,y}, \theta^*) \\ &= p(\rho_{n,y}) \prod_{j=1}^{k} p_{0Y}(\theta_j^*) p(\rho_{n_{j+},x} | \theta_j^*). \end{aligned}$$

Next, using the results of the random partition model of the DP (Antoniak

[1974]), we have

$$p(\rho_n, \theta^*) = \frac{\Gamma(\alpha_y)}{\Gamma(\alpha_y + n)} \alpha_y^k \prod_{j=1}^{k} p_{0Y}(\theta_j^*) \alpha_x(\theta_j^*)^{k_j} \frac{\Gamma(\alpha_x(\theta_j^*))\Gamma(n_{j+})}{\Gamma(\alpha_x(\theta_j^*) + n_{j+})} \prod_{l=1}^{k_j} \Gamma(n_{l|j}).$$

Integrating out θ^* leads to the result. ∎

From Proposition 4.3.1, we gain an understanding of the types of partitions preferred by the EDP and the effect of the parameters. A large value of α_y will encourage more y-clusters, and, given θ^*, a large $\alpha_x(\theta_j^*)$ will encourage more x-clusters within the j^{th} y-cluster. The term $\prod_{j=1}^{k} \prod_{l=1}^{k_j} \Gamma(n_{j,l})$ will encourage asymmetrical (y, x)-clusters, preferring one large cluster and several small clusters, while, given θ^*, the term involving the product of Beta functions contains parts that both encourage and discourage asymmetrical y-clusters. In the special case when $\alpha_x(\theta) = \alpha_x$ for all $\theta \in \Theta$, the random partition model simplifies to

$$p(\rho_n) = \frac{\Gamma(\alpha_y)}{\Gamma(\alpha_y + n)} \alpha_y^k \prod_{j=1}^{k} \alpha_x^{k_j} \frac{\Gamma(\alpha_x)\Gamma(n_{j+})}{\Gamma(\alpha_x + n_{j+})} \prod_{l=1}^{k_j} \Gamma(n_{j,l}).$$

In this case, the overall tendency of term involving the product of Beta functions is to slightly prefer asymmetrical y-clusters with large values of α_x boosting this preference.

As discussed for the DP mixture model, the random partition plays a crucial role, as its posterior distribution affects both inference on the cluster-specific parameters and prediction. For the EDP, it is given by the following proposition.

Proposition 4.3.2 *The posterior of the random partition of the EDP model is*

$$p(\rho_n |\ x_{1:n}, y_{1:n})$$
$$\propto \alpha_y^k \prod_{j=1}^{k} \int_{\Theta} \frac{\Gamma(\alpha_x(\theta))\Gamma(n_{j+})}{\Gamma(\alpha_x(\theta) + n_{j+})} \alpha_x(\theta)^{k_j} dP_{0Y}(\theta)\, g_y(y_j^* | x_j^*) \prod_{l=1}^{k_j} \Gamma(n_{l|j}) g_x(x_{l|j}^*).$$

The proof relies on a simple application of Bayes theorem. In the case of constant $\alpha_x(\theta)$, the expression for the posterior of ρ_n simplifies to

$$p(\rho_n|\ x_{1:n}, y_{1:n}) \propto \alpha_y^k \prod_{j=1}^{k} \frac{\Gamma(\alpha_x)\Gamma(n_{j+})}{\Gamma(\alpha_x + n_{j+})} \alpha_x^{k_j} g_y(y_j^*|x_j^*) \prod_{l=1}^{k_j} \Gamma(n_{l|j}) g_x(x_{l|j}^*).$$

Again, as in (4.4), the marginal likelihood component in the posterior distribution of ρ_n is the product of the cluster specific marginal likelihoods, but now the nested clustering structure of the EDP separates the factors relative to x and $y|x$, being $g(x_{1:n}, y_{1:n}|\rho_n) = \prod_{j=1}^{k} g_y(y_j^*|x_j^*) \prod_{l=1}^{k_j} g_x(x_{l|j}^*)$. Even if the x-likelihood favors many x-clusters, now these can be obtained by sub-partitioning a coarser y-partition, and the number k of y-clusters can be expected to be much smaller than in (4.4).

Further insights into the behavior of the random partition are given by the induced covariate-dependent random partition of the y-parameters given the covariates, which is detailed in the following propositions. We will use the notation $\mathcal{P}_n$ to denote the set of all possible partitions of the first n integers.

Proposition 4.3.3 *The covariate-dependent random partition model induced by the EDP prior is*

$$p(\rho_{n,y}|x_{1:n}) \propto \alpha_y^k \prod_{j=1}^{k} \sum_{\rho_{n_{j+},x} \in \mathcal{P}_{n_{j+}}} \int_\Theta \frac{\Gamma(\alpha_x(\theta))\Gamma(n_{j+})}{\Gamma(\alpha_x(\theta) + n_{j+})} \alpha_x(\theta)^{k_j} dP_{0Y}(\theta) * \prod_{l=1}^{k_j} \Gamma(n_{l|j}) g_x(x_{l|j}^*).$$

Proof. An application of Bayes theorem implies that

$$p(\rho_n|x_{1:n}) \propto \alpha_y^k \prod_{j=1}^{k} \int_\Theta \frac{\Gamma(\alpha_x(\theta))\Gamma(n_{j+})}{\Gamma(\alpha_x(\theta) + n_{j+})} \alpha_x(\theta)^{k_j} dP_{0Y}(\theta) \prod_{l=1}^{k_j} \Gamma(n_{l|j}) g_x(x_{l|j}^*). \tag{4.12}$$

Integrating over $\rho_{n,x}$, or equivalently summing over all $\rho_{n_{j+},x}$ in $\mathcal{P}_{n_{j+},x}$

for $j = 1, \ldots, k$ leads to,

$$p(\rho_{n,y}|x_{1:n}) \propto \sum_{\rho_{n_{1+},x}} \cdots \sum_{\rho_{n_{k+},x}} \alpha_y^k \prod_{j=1}^{k} \int_\Theta \frac{\Gamma(\alpha_x(\theta))\Gamma(n_{j+})}{\Gamma(\alpha_x(\theta)+n_{j+})} \alpha_x(\theta)^{k_j} dP_{0Y}(\theta) \\ * \prod_{l=1}^{k_j} \Gamma(n_{l|j}) g_x(x^*_{l|j}),$$

and, finally, since (4.12) is the product over the j terms, we can pull the sum over $\rho_{n_{j+},x}$ within the product. ∎

This covariate-dependent random partition model will favor y-partitions of the subjects which can be further partitioned into groups with similar covariates, where a partition with many desirable sub-partitions will have higher mass.

Proposition 4.3.4 *The posterior of the random covariate-dependent partition induced from the EDP model is*

$$p(\rho_{n,y}|x_{1:n}, y_{1:n}) \propto \alpha_y^k \prod_{j=1}^{k} g_y(y_j^*|x_j^*) \\ * \sum_{\rho_{n_{j+},x} \in \mathcal{P}_{n_{j+}}} \int_\Theta \frac{\Gamma(\alpha_x(\theta))\Gamma(n_{j+})}{\Gamma(\alpha_x(\theta)+n_{j+})} \alpha_x(\theta)^{k_j} dP_{0Y}(\theta) \prod_{h=1}^{k_j} \Gamma(n_{l|j}) g_x(x^*_{l|j}).$$

The proof is similar in spirit to that of Proposition 4.3.3. Notice the preferred y-partitions will consist of clusters with a similar relationship between y and x, as measured by marginal local model g_y for $y|x$ and similar x behavior, which is measured much more flexibly as a mixture of the previous marginal local models. Again, if $\alpha_x(\theta)$ is constant, the posterior of $\rho_{n,y}$ can be simplified to

$$p(\rho_{n,y}|x_{1:n}, y_{1:n}) \propto \alpha_\theta^k \prod_{j=1}^{k} \frac{\Gamma(\alpha_x)\Gamma(n_{j+})}{\Gamma(\alpha_x+n_{j+})} g_y(y_j^*|x_j^*) \\ * \sum_{\rho_{n_{j+},x} \in \mathcal{P}_{n_{j+}}} \alpha_x^{k_j} \prod_{h=1}^{k_j} \Gamma(n_{l|j}) g_x(x^*_{l|j}).$$

4.3.2 Posterior of the unique parameters

The behavior of the random partition, detailed above, has important implications for the posterior of the unique parameters. Conditionally on the partition, the cluster-specific parameters (θ^*, ψ^*) are still independent, their posterior density being

$$p(\theta^*, \psi^* | y_{1:n}, x_{1:n}, \rho_n) = \prod_{j=1}^{k} p(\theta_j^* | y_j^*, x_j^*) \prod_{l=1}^{k_j} p(\psi_{l|j}^* | x_{j,l}^*),$$

where

$$p(\theta_j^* | y_j^*, x_j^*) \propto p_{0Y}(\theta_j^*) \prod_{i \in S_{j+}} K(y_i; \theta_j^*, x_i),$$
$$p(\psi_{l|j}^* | x_{j,l}^*) \propto p_{0X}(\psi_{l|j}^*) \prod_{i \in S_{j,l}} K(x_i; \psi_{l|j}^*).$$

An important point is that the posterior of θ_j^* can now be updated with much larger sample sizes if the data determines that a coarser y-partition is present. This will result in a more reliable posterior mean, a smaller posterior variance, larger influence of the data compared with the prior.

4.3.3 Covariate-dependent urn scheme

Similar to the DP model, computation of the predictive estimates relies on a covariate-dependent urn scheme, which, given also (θ^*, ψ^*), is

$$s_{y,n+1} | \rho_n, \theta^*, \psi^*, x_{1:n+1} \sim \frac{w_{k+1}^*(x_{n+1})}{c_0} \delta_{k+1} + \sum_{j=1}^{k} \frac{w_j^*(x_{n+1})}{c_0} \delta_j, \quad (4.13)$$

where $c_0 = p(x_{n+1} | \rho_n, \theta^*, \psi^*) * (\alpha_y + n)$ is a normalizing constant,

$$w_{k+1}^*(x_{n+1}) = \alpha_y g_x(x_{n+1}),$$

and for $j = 1, \ldots, k$,

$$w_j^*(x_{n+1}) = \frac{n_{j+} \alpha_y(\theta_j^*)}{\alpha_y(\theta_j^*) + n_{j+}} g_x(x_{n+1}) + \sum_{l=1}^{k_j} \frac{n_{j+} n_{j,l}}{\alpha_y(\theta_j^*) + n_{j+}} K(x_{n+1}; \psi_{l|j}^*).$$

Notice that (4.13) is similar to the covariate-dependent urn scheme of the DP model. The important difference is that the weights, which measure the similarity between x_{n+1} and the j^{th} cluster, are much more flexible.

Under the assumption of constant $\alpha_x(\theta)$ and conjugate P_{0X}, the covariate dependent urn scheme is defined as (4.13) with weights, for $j = 1, \ldots, k$,

$$w_j^{*\prime}(x_{n+1}) = \frac{n_{j+}\alpha_y}{\alpha_y + n_{j+}} g_x(x_{n+1}) + \sum_{l=1}^{k_j} \frac{n_{j+}n_{j,l}}{\alpha_y + n_{j+}} g_x(x_{n+1}|x_{j,l}^*),$$

and normalizing constant $c_0' = p(x_{n+1}|\rho_n, x_{1:n}) * (\alpha_y + n)$.

4.3.4 Prediction

Under the squared error loss function, the prediction of y_{n+1} is

$$\mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}] = \sum_{\mathcal{P}_n \times \mathcal{P}_{n_{j+}}^k} \int_{\Theta^k} \int_{\Psi^{k_+}} [\ldots] dP(\rho_n, \theta^*, \psi^*|y_{1:n}, x_{1:n}), \tag{4.14}$$

$$[\ldots] = \frac{w_{k+1}^*(x_{n+1})}{c_1} E_{G_y}[Y_{n+1}|x_{n+1}] + \sum_{j=1}^{k} \frac{w_j^*(x_{n+1})}{c_1} E_{F_y}[Y_{n+1}|x_{n+1}, \theta_j^*], \tag{4.15}$$

where $c_1 = p(x_{n+1}|y_{1:n}, x_{1:n}) * (\alpha_y + n)$, $\mathcal{P}_{n_{j+}}^k$ represent the product space of $\mathcal{P}_{n_{j+}}$, the set of all partitions of the first n_{j+} integers, over $j = 1, \ldots, k$, and $k_+ = \sum_{j=1}^k k_j$.

The predictive density of y for a new subject with a covariate of x_{n+1} is

$$f(y|y_{1:n}, x_{1:n+1}) = \sum_{\mathcal{P}_n \times \mathcal{P}_{n_{j+}}^k} \int_{\Theta^k} \int_{\Psi^{k_+}} [\ldots] dP(\rho_n, \theta^*, \psi^*|y_{1:n}, x_{1:n}), \tag{4.16}$$

$$[\ldots] = \frac{w_{k+1}^*(x_{n+1})}{c_1} g_y(y|x_{n+1}) + \sum_{j=1}^{k} \frac{w_j^*(x_{n+1})}{c_1} K(y; x_{n+1}, \theta_j^*). \tag{4.17}$$

Similar to the DP model, given the partition, θ^*, and ψ^*, the cluster specific predictive estimates are averaged with covariate-dependent weights, but there are two important differences for the EDP model. The first is that the covariate-dependent weights are defined with a more flexible kernel; in fact, it is a mixture of the original kernels used in the DP model. This means that we have a more flexible measure of similarity in the covariate space. The second difference is that k will be smaller and n_{j+} will be larger with a high posterior probability, leading to a more reliable posterior distribution of θ_j^* due to larger sample sizes and better cluster specific predictive estimates. We will demonstrate the advantage of these two key differences in simulated and applied examples, but first, we discuss sampling procedures.

We note that for example of Section 4.2.4 when $K(y; x, \theta) = N(y; \underline{X}\beta, \sigma^2)$ and the prior for (β, σ^2) is the multivariate normal-inverse gamma with parameters (β_0, C, a_y, b_y), the expressions (4.15) and (4.17) are similar to (4.10) and (4.11) but are defined with the more flexible EDP weights.

4.4 Computations

Inference for the EDP model cannot be obtained analytically and must therefore be approximated. To obtain approximate inference, we rely on Markov Chain Monte Carlo (MCMC) methods and consider an extension of Algorithm 2 of Neal [2000] for the DP mixture model. In this approach, the random probability measure, $\mathbf{P}$, is integrated out, and the model is viewed in terms of $(\rho_n, \theta^*, \psi^*)$. This algorithm requires the use of conjugate base measures P_{0Y} and P_{0X}. To deal with non-conjugate base measures, the approach used in Algorithm 8 of Neal [2000] can be incorporated.

Algorithm 2 is a Gibbs sampler which first samples the cluster label of each subject conditional to the partition of all other subjects, the data, and (θ^*, ψ^*), and then samples (θ^*, ψ^*) given the partition and the data. The first step can be easily performed thanks to the Pólya urn characterization of the DP.

Extending Algorithm 2 for the EDP model is straightforward, since the EDP maintains a simple, analytically computable urn scheme. In particular, letting $s_i = (s_{i,y}, s_{i,x})$ denote the vector containing y-cluster label and x-cluster label for subject i,

$$s_i|\rho_{n-1}^{-i}, \theta^*, \psi^*, x_{1:n}, y_{1:n} \sim \frac{w^*_{k^{-i}+1,1}(y_i, x_i)}{c}\delta_{(k^{-i}+1,1)} + \sum_{j=1}^{k^{-i}} \left(\frac{w^*_{j,k_j^{-i}+1}(y_i, x_i)}{c}\delta_{(j,k_j^{-i}+1)} + \sum_{l=1}^{k_j^{-i}} \frac{w^*_{j,l}(y_i, x_i)}{c}\delta_{(j,l)} \right), \quad (4.18)$$

where for $j = 1, \ldots, k^{-i}$ and $l = 1, \ldots, k_j^{-i}$,

$$w^*_{j,l}(y_i, x_i) = \frac{n_{j+}^{-i} n_{j,l}^{-i}}{\alpha_x(\theta_j^{*-i}) + n_{j+}^{-i}} K(y_i; x_i, \theta_j^{*-i}) K(x_i; \psi_{l|j}^{*-i}),$$

for $j = 1, \ldots, k^{-i}$,

$$w^*_{j,k_j^{-i}+1}(y_i, x_i) = \frac{n_{j+}^{-i} \alpha_x(\theta_j^{*-i})}{\alpha_x(\theta_j^{*-i}) + n_{j+}^{-i}} K(y_i; x_i, \theta_j^{*-i}) g_x(x_i),$$

$$w^*_{k^{-i}+1,1}(y_i, x_i) = \alpha_y g_y(y_i|x_i) g_x(x_i),$$

and

$$c = w^*_{k^{-i}+1,1}(y_i, x_i) + \sum_{j=1}^{k^{-i}} \left(w^*_{j,k_j^{-i}+1}(y_i, x_i) + \sum_{l=1}^{k_j^{-i}} w^*_{j,l}(y_i, x_i) \right).$$

Here, ρ_{n-1}^{-i} represents the partition of the $n-1$ subjects with the i^{th} subject removed where $k^{-i}, k_j^{-i}, n_{j+}^{-i}, n_{j,l}^{-i}$ are defined from ρ_{n-1}^{-i}. Similarly, θ_j^{*-i} and $\psi_{l|j}^{*-i}$ are the unique cluster parameters associated to the clusters of ρ_{n-1}^{-i}.

The algorithm can be summarized as follows:

- For $i = 1, \ldots, n$,
  - if $s_{i,y} = j$ and $n_{j+}^{-i} = 0$,
    - then remove θ_j^* and $\psi_{l|j}^*$ from (θ^*, ψ^*).

- Otherwise, if $s_{i,y} = j$, $s_{i,x} = l$ and $n_{j,l}^{-i} = 0$,
  - then remove $\psi_{l|j}^*$ from ψ^*.
- Next, sample s_i given $\rho_{n-1}^{-i}, \theta^*, \psi^*, x_{1:n}, y_{1:n}$ as defined by equation (4.18).
- If $s_{i,y} = k^{-i} + 1$,
  - sample $\theta_{k^{-i}+1}^*$ given y_i, x_i and $\psi_{1|k^{-i}+1}^*$ given x_i and concatenate them to (θ^*, ψ^*).
- Otherwise, if $s_{i,y} = j$ and $s_{i,x} = k_j^{-i} + 1$,
  - sample $\psi_{k_j^{-i}+1|j}^*$ given x_i and concatenate it to ψ^*.

- For $j = 1, \ldots, k$,
  - sample θ_j^* given (y_j^*, x_j^*), that is from the posterior based on $p_{0Y}(\theta_j^*)$ and $\prod_{i \in S_{j+}} K(y_i; x_i, \theta_j^*)$,
  - and for $l = 1, \ldots, k_j$,
    - sample $\psi_{l|j}^*$ given $x_{j,l}^*$, that is from the posterior based on $p_{0X}(\psi_{l|j}^*)$ and $\prod_{i \in S_{j,l}} K(x_i; \psi_{l|j}^*)$.

The output of the MCMC, $\{\rho_n^s, \psi^{*s}, \theta^{*s}\}_{s=1}^S$, contains approximate samples from the posterior and can be used to estimate the prediction. In particular, the prediction given in equation (4.14) can be approximated by

$$\frac{1}{S}\sum_{s=1}^{S} \frac{w_{k+1}^{*s}(x_{n+1})}{\hat{c}_1} E_{G_y}[Y_{n+1}|x_{n+1}] + \sum_{j=1}^{k^s} \frac{w_j^{*s}(x_{n+1})}{\hat{c}_1} E_{F_y}[Y_{n+1}|x_{n+1}, \theta_j^{*s}],$$

where $w_j^{*s}(x_{n+1})$ for $j = 1, \ldots, k^s + 1$, are as previously defined in (4.13) with $(\rho_n, \psi^*, \theta^*)$ replaced by $(\rho_n^s, \psi^{*s}, \theta^{*s})$ and

$$\hat{c}_1 = \frac{1}{S}\sum_{s=1}^{S} w_{k+1}^{*s}(x_{n+1}) + \sum_{j=1}^{k^s} w_j^{*s}(x_{n+1}).$$

For the predictive density estimate at x_{n+1}, we define a grid of new y values and for each y in the grid, we compute

$$\frac{1}{S}\sum_{s=1}^{S}\frac{w_{k+1}^{*s}(x_{n+1})}{\hat{c}_1}g_y(y|x_{n+1})+\sum_{j=1}^{k^s}\frac{w_j^{*s}(x_{n+1})}{\hat{c}_1}K(y;x_{n+1},\theta_j^{*s}). \quad (4.19)$$

Note that hyperpriors may be included for the precision parameters, α_y and $\alpha_x(\cdot)$, and the parameters of the base measures. For the simulated examples and application, we consider the former. A Gamma hyperprior is assigned to α_y, and $\alpha_x(\theta)$ for $\theta \in \Theta$ are assumed to be i.i.d. from a Gamma hyperprior. At each iteration, α_y^s and $\alpha_x^s(\theta_j^{*s})$ for $j = 1,\ldots,k^s$ are draws from the posterior, which can be sampled using the method described in Escobar and West [1995].

4.5 Simulated example

Here, we consider a toy example that shows the advantages of the EDP, even for moderate values of p. The data was simulated from a mixture of two multivariate normals with $p = 4$, and our aim is to obtain estimates of the regression function and conditional density estimate. We employ the DP mixture model and EDP mixture model as kernel methods to obtain these estimates. A sample size of $n = 200$ was simulated as follows:

$$\begin{aligned} &Y_i|x_i,\beta_i,\sigma_i^2 \overset{ind}{\sim} N(\underline{X}_i\beta_i,\sigma_i^2),\\ &X_i=(X_{1i}\ X_{2i}\ X_{3i}\ X_{4i})'\,|\mu_i,\Sigma_i \overset{iid}{\sim} N_4(\mu_i,\Sigma_i). \end{aligned} \quad (4.20)$$

With probability 1/3,

$$\begin{aligned} &\beta_i=(0\ 0.5\ 0.5\ 0.5\ 0.5)', \quad \sigma_i^2=1/4, \quad (4.21)\\ &\mu_i=\begin{pmatrix}1\\1\\1\\1\end{pmatrix}, \quad \Sigma_i=\begin{pmatrix}1 & 3/4 & 3/4 & 3/4\\ 3/4 & 1 & 3/4 & 3/4\\ 3/4 & 3/4 & 1 & 3/4\\ 3/4 & 3/4 & 3/4 & 1\end{pmatrix}, \end{aligned}$$

and with probability 2/3,

$$\beta_i = (5\ 0.1\ 0.05\ 0.1\ 0)', \quad \sigma_i^2 = 1/4, \tag{4.22}$$
$$\mu_i = \begin{pmatrix} 5 \\ 5 \\ 5 \\ 5 \end{pmatrix}, \quad \Sigma_i = \begin{pmatrix} 1 & 3/4 & 3/4 & 3/4 \\ 3/4 & 1.5 & 1 & 3/4 \\ 3/4 & 1 & 2 & 5/4 \\ 3/4 & 3/4 & 5/4 & 2.5 \end{pmatrix}.$$

We examine the following model for two different choices of Q:

$$\begin{aligned} Y_i|x_i, \beta_i, \sigma_{y,i}^2 &\overset{ind}{\sim} N(\underline{X}_i\beta_i, \sigma_{y,i}^2), \\ X_i|\mu_i, \sigma_{x,i}^2 &\overset{ind}{\sim} \prod_{h=1}^{p} N(\mu_{i,h}, \sigma_{x,h,i}^2), \\ (\beta_i, \sigma_{y,i}^2, \mu_i, \sigma_{x,i}^2)|P &\overset{iid}{\sim} P, \\ \mathbf{P} &\sim Q. \end{aligned}$$

Notice that, as is the practice, we assume independence of X locally.

The first choice of Q is a DP with mass parameter α and base measure $P_{0Y} \times P_{0X}$, where P_{0Y} is the conjugate multivariate normal-inverse gamma prior and P_{0X} is the product of p normal-inverse gamma priors, that is

$$p_{0Y}(\beta, \sigma_y^2) = \mathrm{N}(\beta; \beta_0, \sigma_y^2 C^{-1})\mathrm{IG}(\sigma_y^2; a_y, b_y),$$

and

$$p_{0X}(\mu, \sigma_x^2) = \prod_{h=1}^{p} \mathrm{N}(\mu_h; \mu_{0,h}, \sigma_{x,h}^2 c_h^{-1})\mathrm{IG}(\sigma_{x,h}^2; a_{x,h}, b_{x,h}).$$

The second choice of Q is an EDP with mass parameters α_y and $\alpha_x(\cdot)$ and the same base measure. For both choices, the parameters of the base measure P_{0Y} are

$$\beta_0 = (2.5\ 0.3\ 0.275\ 0.3\ 0.25)', \quad C = \mathrm{diag}\,(0.125\ 12.5\ 12.5\ 12.5\ 12.5)\,;$$
$$a_y = 2, \quad b_y = .25,$$

Table 4.1: Estimated subject-specific regression parameters and the average absolute difference between the estimates and true values for the DP model.

	$\hat{\beta}_{0,i}$	$\hat{\beta}_{1,i}$	$\hat{\beta}_{2,i}$	$\hat{\beta}_{3,i}$	$\hat{\beta}_{4,i}$	$\hat{\sigma^2}_{y,i}$
Subject 2	0.0998	0.3787	0.4472	0.4542	0.3895	0.2844
Subject 4	2.7370	0.2961	0.2914	0.2912	0.2107	0.1845
Subject 5	2.5929	0.2252	0.2797	0.2561	-0.0092	0.2086
Avg. Diff.	1.7576	0.1551	0.1578	0.1138	0.0880	0.0387

and the parameters of the base measure P_{0X} are

$$\mu_0 = (3\ 3\ 3\ 3)', \quad c = (0.75\ 0.75\ 0.75\ 0.75)';$$
$$a_x = (2\ 2\ 2\ 2)', \quad b_x = (1\ 1.25\ 1.5\ 1.75)'.$$

We assign hyperpriors to the mass parameters, where for the first model,

$$\alpha \sim \text{Gamma}(1, 1),$$

and for the second model,

$$\alpha_y \sim \text{Gamma}(1, 1),$$
$$\alpha_x(\beta, \sigma_y^2) \overset{iid}{\sim} \text{Gamma}(1, 1) \quad \forall \beta, \sigma_y^2 \in \mathbb{R}^p \times \mathbb{R}_+.$$

The computational procedures described in Section 4.4 were used to obtain posterior inference with 10,000 iterations and burn in period of 5,000. An examination of the trace plots and autocorrelation plots for the subject specific parameters $(\beta_i, \sigma_{y,i}^2, \mu_i, \sigma_{x,i}^2)$ provided evidence of convergence.

For each subject, we can estimate the subject-specific regression line β_i from the MCMC output:

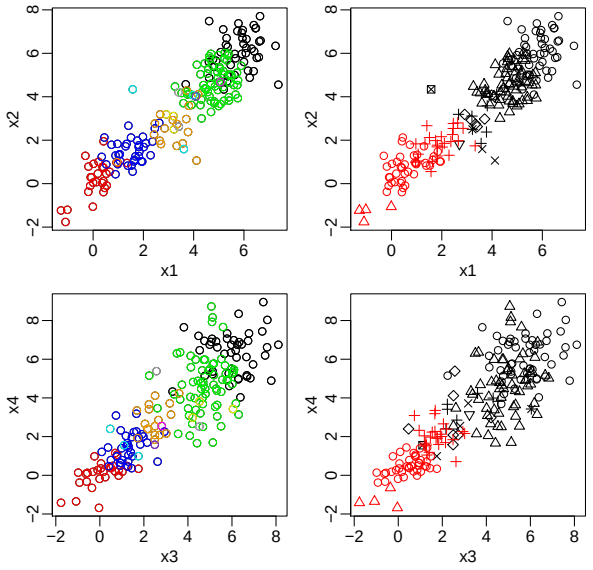
$$\hat{\beta}_i = \frac{1}{S} \sum_{s=1}^{S} \beta_i^s,$$

Table 4.2: Estimated subject-specific regression parameters and the average absolute difference between the estimates and true values for the EDP model.

	$\hat{\beta}_{0,i}$	$\hat{\beta}_{1,i}$	$\hat{\beta}_{2,i}$	$\hat{\beta}_{3,i}$	$\hat{\beta}_{4,i}$	$\hat{\sigma^2}_{y,i}$
Subject 2	0.1573	0.4466	0.5299	0.5177	0.4116	0.2512
Subject 4	3.0939	0.2737	0.2515	0.2522	0.1047	0.2292
Subject 5	4.4113	0.1987	0.1126	0.1267	-0.0764	0.2459
Avg. Diff.	0.5772	0.0909	0.0620	0.0341	0.0846	0.0051

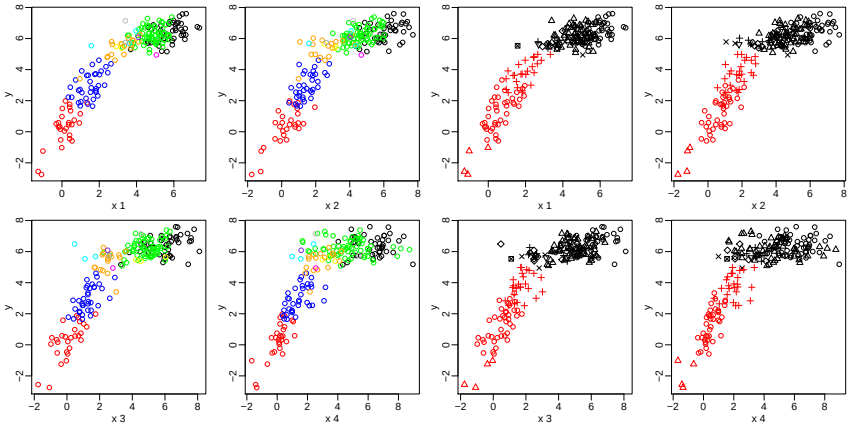
where $\beta_i^s = \beta_j^{*s}$ if $s_i = j$. Since the data is simulated from a mixture of two multivariate normals, we know the true parameters of each subject. Overall, the estimates of the subject-specific parameters are better for the EDP model. This can be seen in Tables 4.1 and 4.2, where we list the estimates of the subject-specific regression lines for three subjects, subjects 2, 4, and 5. The observations of subject 2 were simulated from the first multivariate normal (4.21) and the observations of subjects 4 and 5 were simulated from the second multivariate normal (4.22). The covariates of subject 4, however, can also be reasonably described by the first normal component. Because of this, for both models, the estimated regression line of subject 4 appears to be an average of the regression lines of the two true components (with the EDP putting more weight on the correct group). In Tables 4.1 and 4.2, we also give the average absolute difference between the estimated and true values. Notice that the EDP model gives the lower average absolute differences for all parameters.

Next, we investigate the posterior of the random partition. The posterior of the partition is spread out for both models. This is because many partitions are very similar, differing only in a few subjects, and, thus, many partitions fit the data well (this aspect will be further discussed in the next chapter). We depict a representative partition of DP model in left panels of Figures 4.1 and 4.2 and a representative partition of the EDP model in the right panels. Observations are plotted in the covariate space in Figure 4.1 and in the $x - y$ space in Figure 4.2. For the DP model, observations



(a) DP (b) EDP

Figure 4.1: The partition with the highest estimated posterior probability is plotted in the covariate space. For the DP model, data points are colored according to the partition. For the EDP model, data points are colored according to the y-partition and plotted with different symbols according to the x-partition within each y-cluster.



(a) DP (b) EDP

Figure 4.2: The partition with the highest estimated posterior probability is plotted in the $x - y$ space. For the DP model, data points are colored according to the partition. For the EDP model, data points are colored according to the y-partition and plotted with different symbols according to the x-partition within each y-cluster.

are colored according to the partition, and for the EDP, observations are colored according to the y-partition with different symbols used to depict the x-partition within each y-cluster.

The DP partition depicted in Figures 4.1 and 4.2 is comprised of many clusters. This large number of clusters is caused by the need to approximate the density of X. In fact, the density of $Y|x$ can be recovered with only two kernels, and the y-partition of the EDP depicted in Figures 4.1 and 4.2, with only two y-clusters, is very similar to the true configuration. Indeed, only 3 subjects are placed in the wrong cluster. The (y, x)-partition of the EDP, on the other hand, consists of many clusters and resembles the partition of the DP model.

The posterior of the partition can also be summarized through the posterior of the number of clusters. The DP partitions on average are composed of a large number of clusters, 11.1469, with 89.12% of the partitions comprised of between 8 and 14 clusters. Instead, most of the EDP y-partitions with a positive estimated posterior probability, 29.32%, are composed of only 2 clusters with only a handful of subjects placed in the incorrect cluster, and 77.35% of the partitions are composed of between 2 and 4 y-clusters. The average number of the EDP (y, x)-clusters, similar to the DP, is large, 13.704, with 59.11% of partitions composed of between 11 and 15 clusters.

The posterior estimate of the precision parameter of the DP model is fairly large (2.116), reflecting the high number of clusters present in the partitions with positive posterior mass. The posterior estimate of the y-precision parameter of the EDP model is much smaller (0.5906), while the posterior estimates of $\alpha_x(\cdot)$ range between 0.5 and 2. Figure (4.3) displays posterior estimates of $\alpha_x(\cdot)$ as a function of the parameters. For high values of the intercept and small values of the slopes, which is characteristic of second model used in simulations (4.22), the posterior estimate of $\alpha_x(\cdot)$ is higher. This means that we need more kernels to approximate the density of x in the second component (4.22). The variance, σ^2, appears to be uninformative for $\alpha_x(\cdot)$. This is due to the fact that σ^2 is the same for both of the components used in simulations.

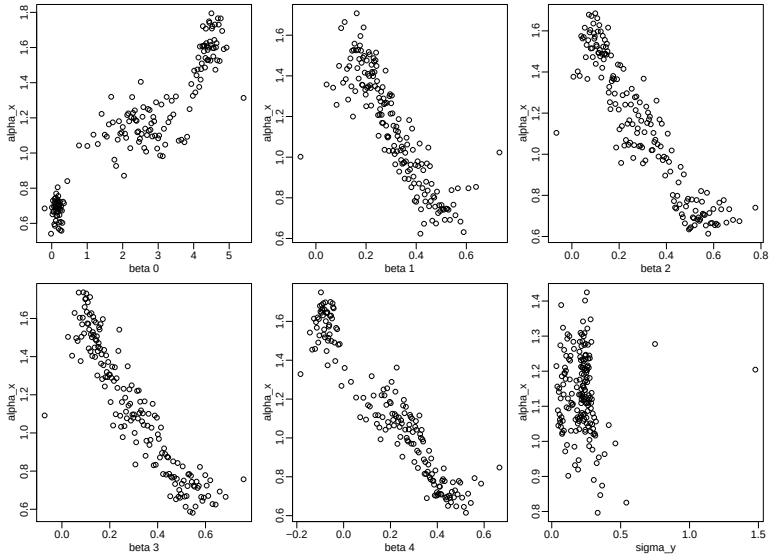


Figure 4.3: Posterior estimates of $\alpha_x(\cdot)$ for different values of β and σ^2.

Both the DP and EDP models are likely to be consistent, that is, as the sample size goes to infinity, the estimates of the regression function and conditional densities will be close to the truth. However, in practice, sample sizes are finite, and consistency properties, while desirable, may hide what happens in finite samples. Thus, the desirable model would be the one that leads to more efficient estimators, in terms of smaller estimation errors and less variability. In Section 4.3, we discussed the increased efficiency of the EDP model. Here, we simulate $m = 100$ new covariates from (4.20) and compute the true regression function $\mathrm{E}[Y_{n+j}|x_{n+j}]$ and conditional density $f(y|x_{n+j})$ for each new subject. To quantify the gain in efficiency of the EDP model for our simulated example, we calculate the the prediction and predictive density estimates from both models and compare them with the truth.

Judging from both the empirical l_1 and l_2 prediction errors, the EDP model outperforms the DP model, although the improvement is not dras-

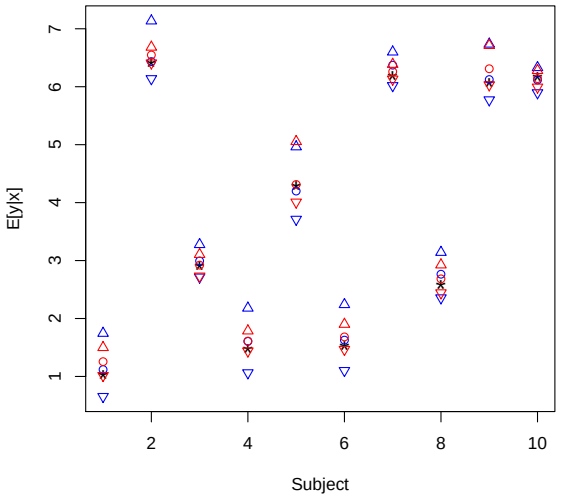


Figure 4.4: The prediction of the response is plotted against subject index for the first 10 new subjects, where the prediction is represented with circles (blue for the DP and red for the EDP) with the true prediction (as black stars). The credible intervals are depicted using triangles (blue for the DP and red for the EDP).

tic. In particular, the l_1 prediction errors for the DP and EDP model respectively are 0.1258 and 0.1107, and the l_2 prediction errors are 0.1641 and 0.1405. The comparison of the credible intervals is more interesting. The larger cluster sample sizes allow for tighter credible intervals, almost uniformly in x, and a quite impressive tightening in some cases.

Due to the multivariate nature of x, visualization of the regression function and credible intervals is difficult. In an attempt at visualization, we have provided a plot (Figure 4.4) displaying the prediction against subject index for the first 10 subjects. The true prediction is denoted by a black star, the estimated prediction is denoted by a circle (blue for

Table 4.3: Estimated prediction with the lower and upper 95% credible bounds for the first 5 new subjects for the DP and EDP models.

Subject	1	2	3	4	5
$\mathrm{E}[y\|x]$	1.063	6.437	2.933	1.506	4.323
$\widehat{\mathrm{E}}_{\mathrm{DP}}[y\|x]$	1.119	6.434	2.994	1.605	4.200
$\widehat{\mathrm{E}}_{\mathrm{EDP}}[y\|x]$	1.256	6.547	2.921	1.611	4.313
$\widehat{l}_{\mathrm{DP}}(x)$	0.654	6.138	2.715	1.063	3.712
$\widehat{l}_{\mathrm{EDP}}(x)$	1.016	6.410	2.732	1.439	4.008
$\widehat{u}_{\mathrm{DP}}(x)$	1.745	7.136	3.277	2.180	4.963
$\widehat{u}_{\mathrm{EDP}}(x)$	1.499	6.683	3.106	1.786	5.055

Table 4.4: Estimated prediction with the lower and upper 95% credible bounds for the following 5 subjects for the DP and EDP models.

Subject	6	7	8	9	10
$\mathrm{E}[y\|x]$	1.561	6.217	2.615	6.102	6.199
$\widehat{\mathrm{E}}_{\mathrm{DP}}[y\|x]$	1.627	6.372	2.765	6.124	6.116
$\widehat{\mathrm{E}}_{\mathrm{EDP}}[y\|x]$	1.683	6.260	2.684	6.310	6.140
$\widehat{l}_{\mathrm{DP}}(x)$	1.102	6.018	2.356	5.773	5.895
$\widehat{l}_{\mathrm{EDP}}(x)$	1.465	6.130	2.443	6.032	5.993
$\widehat{u}_{\mathrm{DP}}(x)$	2.241	6.600	3.142	6.736	6.330
$\widehat{u}_{\mathrm{EDP}}(x)$	1.901	6.387	2.925	6.713	6.280

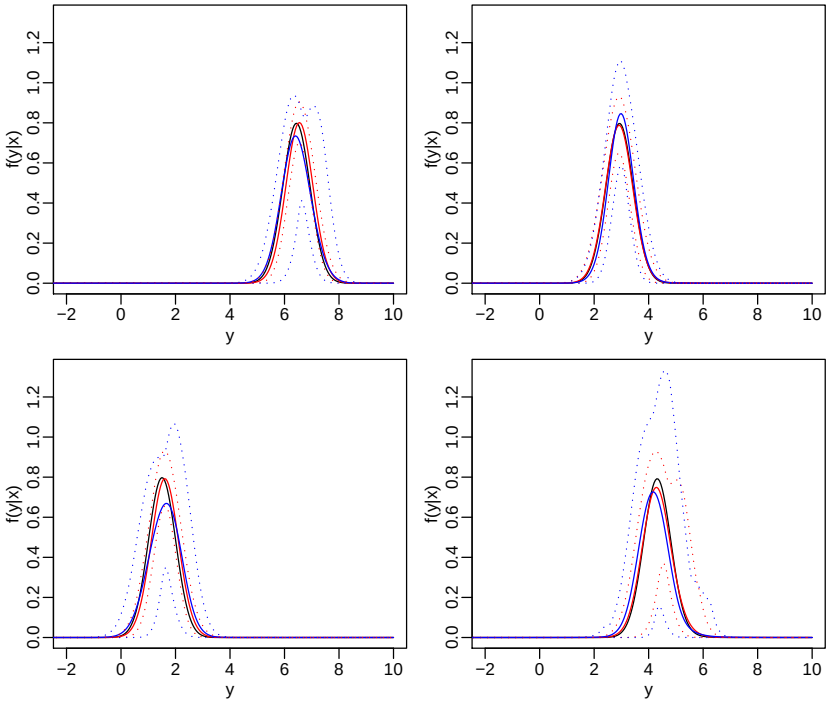


Figure 4.5: The predictive density estimates (blue for the DP and red for the EDP) for 4 new covariate values with the true conditional density in black. The point-wise 95% credible bounds are also displayed in blue dashed lines for the DP and red dashed lines for the EDP.

the DP and red for the EDP), and the lower and upper credible bounds are denoted by triangles (blue for the DP and red for the EDP). The important thing to take away from this plot is the unnecessarily wide credible intervals depicted by the blue triangles. These estimates are also listed in Tables 4.3 and 4.4, with the true prediction in the first column, the estimated prediction for both the DP and EDP model in the second and third columns, and the lower and upper 95% credible bounds in the last columns.

The predictive density estimate for all new subjects was also computed by evaluating (4.19) at a grid of y-values. To evaluate the performance of the models, we computed the empirical l_1 distance between the true and estimated conditional densities for each of the new covariate values. Again the EDP model outperforms the DP model with an average of l_1 distance of 0.1859 for the EDP versus 0.2502 for the DP, a maximum l_1 distance of 0.5673 against 0.7589, and a minimum l_1 distance of 0.00996 against 0.03817. Again, this conclusion becomes more dramatic when comparing the pointwise credible intervals. Figure 4.5 displays the true conditional density in black for four new covariate values with the estimated conditional densities in blue for the DP and red for the EDP. The pointwise 95% credible intervals are shown as dashed lines (blue for the DP and red for the EDP). For most subjects, the estimated conditional densities of the DP model tend to be flatter (as is the case in the plot at the bottom left hand corner of Figure 4.5). However, for some subjects the DP model overestimates the density at the mode (see the plot at the top right hand corner of Figure 4.5). The pointwise 95% credible intervals are almost uniformly wider both in y and x for the DP model, sometimes drastically so. In fact, for many new covariates the flatter estimate of the DP model resembles the lower 95% credible intervals of the EDP model around the mode. It is important to note that while the credible intervals of the EDP model are considerably tighter, they still contain the true density.

4.6 Alzheimer's disease study

The first attempts to automatically diagnose Alzheimer's disease based on neuroimages focused on regions of the brain known to be affected by the disease, called regions of interest (ROI). For each patient, the volume of the ROI is calculated, and this volume is compared between groups using parametric methods such as linear discriminant analysis or logistic regression. This approach has had some successful results with estimated accuracy rates ranging from 70% up to 90% (Davatzikos et al. [2008b],Wolf et al. [2001], Laakso et al. [1998]), depending on ROI used and the severity

of the disease for the observed subjects.

More recent approaches have attempted to predict disease status based on the entire brain image, in order to capture the complex pattern of atrophy associated with AD. While these methods have had some successful results (Davatzikos et al. [2008a], Davatzikos et al. [2008b], Kloppel et al. [2008]), the massive dimension and complexity of the data introduce serious challenges.

Although a whole brain analysis allows for the possibility to capture the heterogeneous pattern of atrophy across and within brain regions, it relies on the tissue density at single voxel, a quantity which is not reliable or interpretable. On the other hand, the volume of a ROI is reliable and easily interpreted, but one can not capture the heterogeneous pattern of atrophy within the region.

An alternative option between these two extremes is to diagnose patients based on a large number of ROIs and subregions of ROIs. In this direction, we examine the diagnostic ability of $p = 15$ structures using Bayesian nonparametric methods. Nonparametric techniques are needed to capture complex interactions, and the Bayesian prior provides a built-in mechanism for shrinkage and inclusion of prior information about the relationship between the ROIs and the disease. In particular, we consider the models discussed in Section 4.2 and 4.3.

The ADNI dataset analysed here consists of summaries of fifteen brain structures computed from the structural Magnetic Resonance image obtained at the first visit for 377 patients, of which 159 have been diagnosed with AD and 218 are cognitively normal (CN). The covariates include whole brain volume (BV), intracranial volume (ICV), volume of the ventricles (VV), left and right hippocampal volume (LHV, RHV), volume of the left and right inferior lateral ventricle (LILV, RILV), thickness of the left and right middle temporal cortex (LMT, RMT), thickness of the left and right inferior temporal cortex (LIT, RIT), thickness of the left and right fusiform cortex (LF, RF), and thickness of the left and right entorhinal cortex (LE, RE). Volume is measured in cm^3 and cortical thickness is measured in mm.

AD is associated with a loss of white and grey matter and an increase in cerebrospinal fluid with a pattern of tissue loss and fluid gain that is spatially distributed over many regions. Whole brain volume measures the total volume of white and grey matter. Thus, we expect AD patients to have smaller brain volumes compared to cognitively normal patients. Similarly, since the ventricles is a set of structures containing cerebrospinal fluid, we expect AD patients to have larger ventricular volume. Total intracranial volume measures the volume in the cranium, including volume of grey matter, white matter, and cerebrospinal fluid. It is determined during childhood, and doesn't decrease with age or disease, therefore AD patients should have smaller brain to intracranial volume ratios and larger ventricular to intracranial volume ratios. However, this relationship has been contested in literature with some studies finding that larger intracranial volume may protect against AD while other studies have negated this finding (see Jenkins et al. [2005]).

The left and right hippocampi are composed of grey matter and located at the base of the brain. Hippocampal volume is the most common ROI used in studies because it is relatively easy to identify and known to be affected by the disease. In particular, loss of hippocampal volume is characteristic of AD, and some studies have also found evidence of asymmetrical tissue loss between the left and right hippocampi in AD patients (Shi et al. [2009]). The inferior lateral ventricles are part of the ventricles and are known to increase with AD. They are located adjacent to the medial temporal lobe structures, which experience tissue loss in early stages of AD, and therefore, may exhibit faster rates of volume increase compared with the entire ventricular volume, especially during early stages of the disease.

The cerebral cortex is the outer layer of brain tissue and is composed of grey matter. Cortical thickness measures the thickness of the cerebral cortex by calculating the local distance between the white matter/grey matter boundary and the grey matter/cerebrospinal fluid boundary and averaging these local distances across the entire cortex or regions within the cortex, in this case, the middle temporal cortex, inferior temporal

cortex, fusiform cortex, and entorhinal cortex. The regions used here are all located in the temporal lobe, a region known to be affected by AD. Lerch et al. [2005] had some successful results classifying patients based on the cortical thickness of twenty-five different regions, particularly with the entorhinal cortex, but also found evidence of heterogeneity of the thickness within region.

The response is a binary variable with 1 indicating a cognitively normal subject and 0 indicating a subject who has been diagnosed with AD. The covariate is the 15-dimensional vector of measurements of various brain structures. Our model builds on local probit models and can be stated as follows:

$$\begin{aligned} Y_i|x_i, \beta_i &\overset{ind}{\sim} \text{Bern}(\Phi(\underline{X}_i\beta_i)), \\ X_i|\mu_i, \sigma_i^2 &\overset{ind}{\sim} \prod_{h=1}^{p} N(\mu_{i,h}, \sigma_{i,h}^2), \\ (\beta_i, \mu_i, \sigma_i^2)|P &\overset{iid}{\sim} P, \\ \mathbf{P} &\sim Q. \end{aligned}$$

The analysis is first carried using a DP prior for $\mathbf{P}$ with mass parameter α and base measure $P_{0Y} \times P_{0X}$, with

$$P_{0Y} = N(0_p, C^{-1}),$$

where C^{-1} is a diagonal matrix with diagonal elements

$$(400, .0001, .0001, 0.0004, 4, 4, .25, .25, 4, 4, 4, 4, 1, 1, 1, 1),$$

and

$$P_{0X} = \prod_{h=1}^{p} NIG(\mu_{0,h}, c_{x,h}, a_{x,h}, b_{x,h}),$$

where

$$\mu_0 = (1000, 1450, 45, 3.25, 3.25, 2, 2, 2.4, 2.4, 2.5, 2.5, 2.3, 2.3, 2.75, 2.75)',$$

$$c_{x,h} = 1/2, a_{x,h} = 2 \; \forall h,$$

$$b_x = (10000, 10000, 150, .25, .25, .25, .25, .04, .04, .04, .04, .04, .04, .1, .1)'.$$

The mass parameter is given a hyperprior of

$$\alpha \sim \text{Gamma}(1, 1).$$

We chose to center the base measure for β on zero because even though we have prior belief about how each structure is related to AD individually, the joint relationship may be more complex. For simplicity, the covariance matrix is diagonal. The variances were chosen to reflect belief in the maximum range of the coefficient for each brain structure. We also explored the idea of defining C through a g-prior, where $C^{-1} = g(\underline{X}'\underline{X})^{-1}$ with g fixed or given a hyperprior. However, this proposal was unsatisfactory because prior information about the maximum range of the coefficient for each brain structure is condensed in a single parameter g. For example, there was no way to incorporate the belief that while the variability of hippocampal volume and inferior lateral ventricular volume are similar, the correlation between hippocampal volume and disease status is stronger.

The parameters of the base measure for X where chosen based on prior knowledge and exploratory analysis of the average volume and cortical thickness of the brain structures (μ_0) and variability (b_x). The parameter a_x was chosen to equal 2, so that mean of the inverse gamma prior is properly defined and the variance is relatively large. The parameter c_x is equal to 1/2 to increase variability of μ given σ_x.

In this example, correlation between the measurements of the brain structures is expected. However, for statistical and computational reasons, we assume local independence of the covariates within kernel. Due to this local independence assumption as well as the non-normal behavior present in the univariate histograms of the covariates, we expect many kernels will be needed to approximate the density of X. The conditional density of the response, on the other hand, may not be so complicated. This motivates the choice of an EDP prior with the same base measure $P_{0Y} \times P_{0X}$ and mass parameters α_y and $\alpha_x(\cdot)$. Again, the mass parameters are assigned

Table 4.5: Estimated subject-specific slopes of brain volume, intracranial volume, ventricular volume, left hippocampal volume, and right hippocampal volume for the DP model.

Subj.	BV Slope	ICV Slope	VV Slope	LHV Slope	RHV Slope
1	0.0013	-0.0014	-0.0117	0.5683	1.1572
2	-0.0003	-0.0010	-0.0007	-0.0811	-0.3907
3	-0.0024	-0.0040	0.0049	1.2305	0.3171
4	-0.0032	-0.0047	0.0046	1.3197	0.4385

hyperpriors of

$$\alpha_y \sim \text{Gamma}(1,1),$$
$$\alpha_x(\beta) \overset{iid}{\sim} \text{Gamma}(1,1) \quad \forall \beta \in \mathbb{R}^{p+1}.$$

As discussed in Section 3.4.2, if $\alpha_x(\beta) \approx 0$ for all $\beta \in \mathbb{R}^{p+1}$ the model converges a DP mixture model, suggesting that the extra flexibility of the EDP is unnecessary. On the other hand, $\alpha_y \approx 0$ suggests that a linear model is sufficient for modelling the conditional response distribution.

The data were randomly split into a training sample of size 185 and a test sample of size 192. Inference for observed sample of 185 patients is based on the algorithm explained in the Section 4.4 with the added step of sampling a latent normal variable to deal with the binary response. For both results the number of iterations is 30,000 with burn in period of 10,000. From an examination of the trace and autocorrelation plots for the subject specific parameters $(\beta_i, \mu_i, \sigma_i^2)$, convergence appears to be reached.

Tables 4.5 and 4.6 list the estimated slopes of brain volume, intracranial volume, ventricular volume, left hippocampal volume, and right hippocampal volume for the first four subjects. Notice that the results differ both across subjects, suggesting that a nonparametric approach may be necessary, and across models, suggesting that the added flexibility of the EDP may be useful for this dataset.

The DP based model requires many kernels to approximate the joint

Table 4.6: Estimated subject-specific slopes of brain volume, intracranial volume, ventricular volume, left hippocampal volume, and right hippocampal volume for the EDP model.

Subj.	BV Slope	ICV Slope	VV Slope	LHV Slope	RHV Slope
1	-0.0005	0.0001	-0.0056	0.9312	0.0494
2	-0.0051	-0.0042	-0.0011	-0.1999	-0.4315
3	-0.0073	0.0008	-0.0009	2.0326	0.0482
4	-0.0071	-0.0003	-0.0009	2.1518	0.1862

distribution. The average number of kernels is 16.035, the mode is 16 (34.13%), and with a high probability (85.13%), the number of kernels falls between 15 and 17. This high number of kernels is mostly driven by the need to obtain a good approximation to the marginal density of the high-dimensional X. The EDP allows a coarser y-partition for the conditional density of $Y|x$, and the estimated number of y-kernels is much less. The average number of y-kernels is 3.6824, the mode is 3 (78.1%), and with an estimated 96.97%, the number of y-kernels falls between 3 and 4.

The estimated precision parameter of the DP based model is large, 3.2954, while the estimated y-precision parameter of the EDP based model is much smaller, 0.54. This again, reflects the fact that the many kernels required by the DP based model are need to approximate the density of X. The estimated values of the x-precision parameters for various values of β are depicted in Figure 4.6. Values of β closer to the average are associated with higher estimated values of $\alpha_x(\beta)$. This means that y-clusters with average values of β need more kernels for the density of X than the y-clusters with more extreme values of β. In fact, the y-partitions with a positive estimated posterior probability generally consist of one large cluster and a few small clusters. The large group has more average values of β, but is heterogeneous in x, while the smaller groups tend to have more extreme values of β, but are fairly homogeneous in x.

The posterior of the partition is fairly flat for the DP and EDP models.

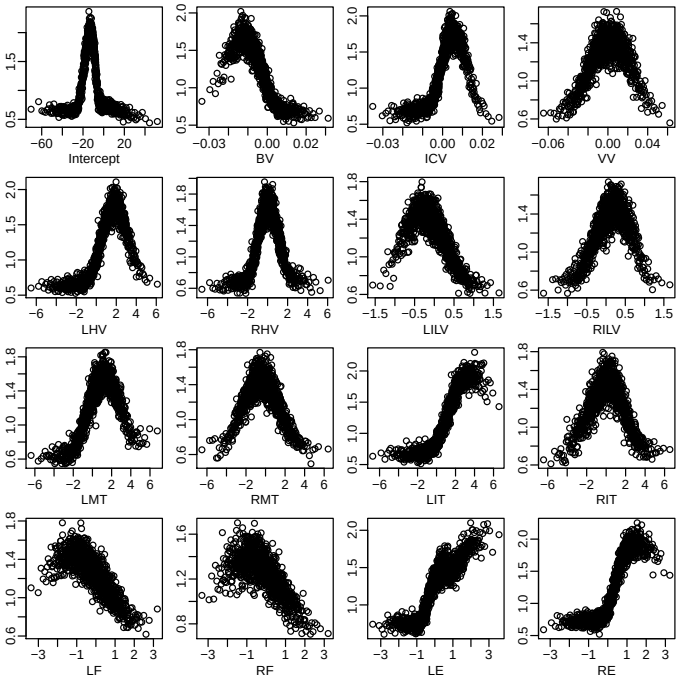


Figure 4.6: Estimated x precision parameters as a function of β_i for $i = 0, 1, ..., p$.

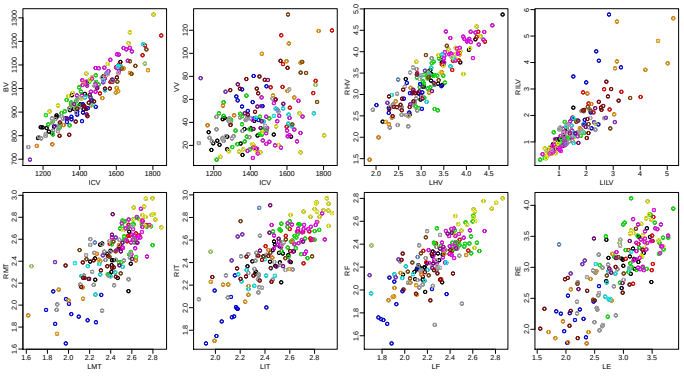


Figure 4.7: Data points are plotted in the covariate space and colored by the partition with the highest posterior probability for the DP model.

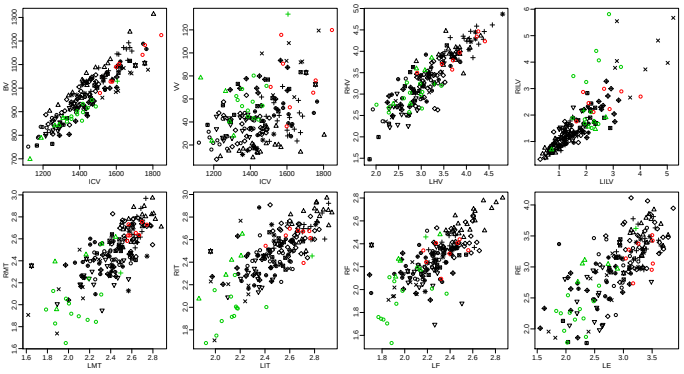


Figure 4.8: Data points are plotted in the covariate space and colored by the y-partition with the highest posterior probability for the EDP model. The plot includes symbols representing an x-partition within each y-cluster.

Table 4.7: DP: Estimated probability of being healthy for 10 subjects with upper and lower 95% credible bounds.

Healthy	Predicted Prob.	Lower Bound	Upper Bound
0	0.4908	0.0001	1
1	0.829	0.1833	1
1	0.5944	0.0659	0.9902
1	0.9653	0.5597	1
1	0.6424	0	1
1	0.8891	0.4751	0.9998
0	0.2971	0	0.998
1	0.9944	0.9301	1
1	0.9866	0.8712	1
1	0.8771	0.431	1

Again, this is due to the fact that there are many similar partitions which fit the data well. A representative partition, the partition with the highest estimated posterior probability, for the DP mixture model is depicted in Figure 4.7, where the data points are plotted in the covariate space and colored by the partition. Notice the high number of kernels with small sample sizes within each cluster. Figure 4.8 depicts a representative partition, the partition with the highest estimated posterior probability, for the EDP mixture model, where the data points are plotted in the covariate space and colored by the y-partition with different symbols for the x-partition within each y-cluster. Sample sizes within kernel are larger, especially for the black cluster.

To quantify the gain in efficiency with the EDP model, we estimated the predictive probability of being healthy for the subjects in the test set. Under the 0-1 loss function, subjects are diagnosed with the disease if the predicted probability of being healthy is less than 0.5. The DP model has an accuracy of 82.8125%, with 159 of the 192 subjects correctly classified. The EDP model does better; 168 subjects are correctly classified, resulting in an accuracy of 87.5%. This is due to the increased sample sizes within

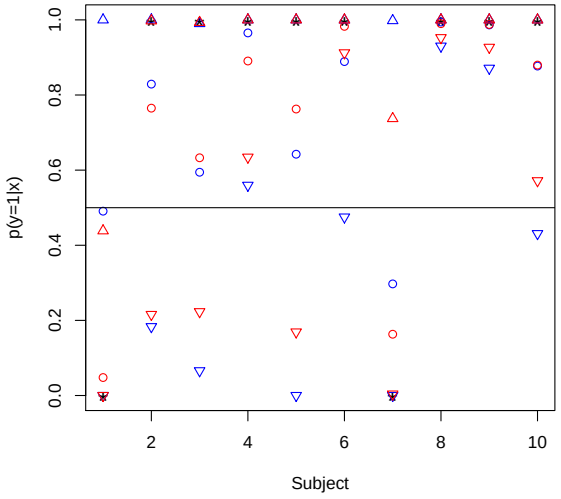


Figure 4.9: Plots the predicted probability of being healthy against subject index for 10 new subjects, where the prediction is represented with circles (blue for the DP and red for the EDP) with the true outcome (as black stars). The credible intervals are depicted using triangles (blue for the DP and red for the EDP).

Table 4.8: EDP: Estimated probability of being healthy for 10 subjects with upper and lower 95% credible bounds.

Healthy	Predicted Prob.	Lower Bound	Upper Bound
0	0.0479	0	0.4384
1	0.765	0.2157	0.9973
1	0.6329	0.2231	0.9932
1	0.8906	0.6346	1
1	0.7627	0.1691	0.9999
1	0.9828	0.9123	0.9998
0	0.1632	0.0039	0.7373
1	0.9899	0.9526	0.9999
1	0.9877	0.927	1
1	0.8797	0.5718	0.9998

cluster leading to more reliable posterior inference within cluster.

A very interesting aspect of the results is found from comparing the credible intervals of the predicted probability of being healthy for the new subjects. By allowing for a coarser y-partition when appropriate, the increased cluster sample sizes of the EDP model allow for much tighter credible intervals. This is shown in Tables 4.7 and 4.8 which give the predicted probability of being healthy for 10 subjects along with lower and upper bounds for 95% credible intervals. These results are also displayed graphically in Figure 4.9. Notice the tighter credible intervals for the EDP model with some dramatic examples given by subjects 1 and 6. In fact, if we consider the number subjects correctly classified with at least 95% probability, this number is much higher for the EDP model, 103 (66 healthy subjects and 37 sick subjects), than for the DP model, 80 (48 healthy subjects and 32 sick subjects). Yet, the number of subjects that are incorrectly classified with at least 95% probability is the same (6) for both models. This is particularly important for the AD example because not only are more subjects correctly diagnosed, but confidence in the diagnosis is higher for the EDP model.

For the EDP, most y-partitions consist of three clusters. There is one large cluster composed of subjects with an volume and cortical thickness close to the overall average and high variability but few extreme values. In this group, the relationship between the brain structures and disease status reflects prior belief, and a small left hippocampal volume and a thin left inferior temporal cortex particularly increase the probability of the disease. The two smaller clusters consist of subjects with extreme x values of large brain tissue volumes and cortical thickness for the first and small brain tissue volumes and cortical thickness for the second. Interestingly, the first group has high intracranial volume, while the second group has low intracranial volume and also displays lower brain volumes relative to intracranial volume. Both groups have high ventricular volume, and the second group has particularly thin cortical structures. Subjects in the first group are mostly classified as healthy with a high probability, but higher ventricular volume and lower brain tissue volume and cortical thickness will decrease this probability, although the change is gradual. The second group is classified as sick with a high probability.

As discussed in the beginning of the section, the DP model and the generalized linear regression model are special cases of the EDP model. The results of EDP model imply that DP model is not appropriate for this data and, in fact, the predictive performance is worse under the DP model. However, the small posterior estimate of α_y suggests that a generalized linear model may be sufficient for this dataset. In fact, the accuracy of prediction for the new subjects is not much worse for the generalized linear regression model. The results depend on the choice of the link function; for most choices, 162 subjects are correctly classified with an accuracy rate of 84.375%, but with a probit link function, this number increases to 165 with an accuracy of 85.9375%. The generalized linear model does, as expected, give tighter credible intervals for some individuals, but at the expense of a slightly smaller number of individuals correctly classified.

To compare the predictive results of EDP model with other nonparametric techniques, we consider support vector machines, Gaussian processes, and random forests, which are implemented in the *kernlab* and

randomForest packages in **R**. Depending on the kernel choice, the results with support vector machines range between 162 to 166 subjects correctly classified (84.375%- 86.4583% accuracy rate), and for Gaussian processes, the range is 163 to 167 subjects correctly classified (84.8958%- 86.9792% accuracy rate) with the best results for the squared exponential kernel and polynomial kernel function, respectively. The best results are obtained with random forests, where, as for the EDP model, 168 subjects are correctly classified. Thus, the predictive results of the EDP model are comparable with, if not better than, other standard nonparametric classification methods.

4.7 Discussion

In this chapter, we have highlighted a drawback of DP mixture models when the aim is estimation of the regression function and conditional density. We have proposed a simple, but efficient, solution based on the EDP, which overcomes the problems of the DP mixture model by introducing a nested partition structure. An important feature of the proposed EDP mixture model is that computations remain relatively simple. To provide formal validation of the EDP mixture model, a direction of further research includes the study of theoretical properties.

In Bayesian nonparametric literature, the standard step is to study posterior consistency. Consistency results for the regression function and conditional density estimates of the DP mixture model are likely to hold for a large class of data generating densities. To prove such results, one would first establish consistency of the joint density estimate and then study the implications for the regression function and conditional density.

The literature on consistency for a random density constructed through a DP mixture model is substantial. To be useful here, available results would need to be extended to allow a more general multivariate kernel. Initial work focused on univariate location mixtures (Ghosal et al. [1999]), and subsequent work considered univariate location-scale mixtures (Ghosal and van der Vaart [2001], Tokdar [2006]), multivariate location-

scale mixtures with a single scale parameter (Wu and Ghosal [2008], Tokdar [2011]), and multivariate location mixtures with a general covariance matrix (Wu and Ghosal [2010]). Our interest is in multivariate location-scale mixtures where the joint kernel is parametrized in terms of the parameters of the univariate conditional and the multivariate marginal with the further assumption that the marginal is the product of p location-scale kernels.

Extending the available consistency results to the DP mixtures of interest should not be too difficult. Some initial work is given in Hannah et al. [2011], where weak consistency of the joint density estimate is studied and asymptotic unbiasedness of the regression function is shown to follow under mild conditions. However, the data generating density is restricted to have compact support and the covariate is assumed to be one-dimensional. We have started to examine weak consistency of the joint density for the DP model studied here with multivariate covariates under milder conditions on the data generating density, but, as the work is still under development, we will not discuss it here.

Furthermore, since weak consistency in Bayesian nonparametric mixture models relies on weak consistency of the random mixing measure, weak consistency is also likely to hold for the proposed EDP mixture model. Strong consistency can also be expected, although it may be more difficult to prove, since most results use properties of the DP in the proof.

However, these consistency results disguise what happens in finite samples. In this chapter, we have shed light on issues of the DP mixture model that can arise in finite samples for moderate to large values of p. Through careful examination of the prediction and predictive density, we have shown that the proposed EDP mixture model can lead to more efficient estimates, in terms of smaller estimation errors and tighter credible intervals.

To quantify this efficiency, we studied two examples, one simulated and one based on real data. In future work, we aim to develop theoretical properties to measure this gain in efficiency based on finite samples. As a starting point, we have reviewed literature on predictive model comparison

(San Martini and Spezzaferri [1984], Laud and Ibrahim [1995], Gelfand and Ghosh [1998]), but would also like to examine finite sample bounds on the probability that regression function or conditional density is contained within some interval of the truth.

Finally, for the AD study, we would like to stress the importance of the predictive improvements of the EDP mixture model over the DP mixture model in this example; not only does the EDP model lead to an improvement in diagnostic accuracy, but it also provides higher credibility in the diagnosis. In a further comparison with other standard nonparametric methods, the EDP mixture model performed just as good, if not better.

We should also mention that the generalized linear model is special case of the EDP mixture model; thus, (with a hyperprior on the precision parameters) the model is able to recognize if the simpler generalized linear model is sufficient for the data. For the brain structures included in the study, the results provided weak evidence for the EDP mixture model over the generalized linear model, and in fact, the predictive performance is slightly improved. Furthermore, we expect that with additional covariates the model will become more advantageous as more complex interaction terms are expected. In future work, we would like to expand the analysis to include the volume and cortical thickness of other structures or possibly (a subset of) the entire image as well as summaries based on other types of neuroimages. A potential downfall is that computations may become heavy with increasing p due to the large number of x-kernels. In that case, we could consider more flexible kernels for x, but that would necessarily increase the number of parameters within each x-kernel. Simulation studies would be needed to examine the trade-off between the number of x-kernels and the number of parameters within each x-kernel.

Chapter 5

Restricted Dirichlet process mixtures

This chapter examines the predictive performance of Bayesian nonparametric mixture models for regression, focusing on the regression function. The random partition plays a crucial role in the prediction, and in regression settings, it is often reasonable to assume that this partition depends on the proximity of the covariates. Models with constant weights do not incorporate this knowledge, and we find that these models can perform quite poorly. Models with covariate-dependent weights encourage covariate-proximity based partitions, which can lead to remarkably improved prediction. However, closer examination of the random partition yields further complications, which arise due to the huge number of total partitions. To overcome this, we propose to modify the probability law of the random partition to strictly enforce the notion of covariate proximity, while still maintaining certain properties of the DP. This allows the distribution of the partition to depend on the covariate in a simple manner and greatly reduces the total number of possible partitions, resulting in improved prediction and faster computations. Numerical illustrations will be presented.

This chapter contains joint work with Stephen G. Walker and Sonia

Petrone and is based on Wade et al. [2012].

5.1 Introduction

Flexible estimation of the regression function is an important research problem. The literature is vast including Breiman et al. [1984], Hastie and Tibshirani [1990], Friedman [1991], Neal [1996], Denison et al. [2002], Vidakovic [2009], and Rasmussen and Williams [2006]. In these proposals, the basic model is of type

$$Y_i = m(x_i) + \sigma\,\varepsilon_i, \tag{5.1}$$

where $m(\cdot)$ is the flexible regression function and the errors have a simple i.i.d. standard normal distribution.

Bayesian nonparametric mixture models for regression have an important advantage over models of type (5.1) in that they significantly relax the assumptions on the error distribution. In particular, the errors may evolve flexibly with x, but the regression function still maintains a flexible structure. In this chapter, our general aim is to examine in detail the predictive performance of Bayesian nonparametric mixture models for flexible estimation of the regression function.

Before proceeding, we would like to underline that, under the quadratic loss function, the estimated regression function, $\hat{m}(\cdot)$ at a new covariate value of x_{n+1}, which is

$$\hat{m}(x_{n+1}) = \mathrm{E}[m(x_{n+1})|y_{1:n}, x_{1:n+1}],$$

is equivalent to the prediction of the response at x_{n+1}, which is

$$\hat{Y}(x_{n+1}) = \mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}].$$

Thus, properties of estimated regression function correspond to properties of the prediction.

In this chapter, we will assume the response is univariate and continuous. The general form of the Bayesian nonparametric mixture model that

we will study is

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j(x)\mathrm{N}(y; \tilde{\mu}_j(x), \tilde{\sigma}_j^2(x)), \tag{5.2}$$

where P_x is a realization of

$$\mathbf{P}_x = \sum_{j=1}^{\infty} w_j(x)\delta_{(\tilde{\mu}_j(x),\tilde{\sigma}_j^2(x))}.$$

Model (5.2) implies that the choice of $m(\cdot)$ is given by

$$m(x) = \mathrm{E}[Y \mid x, P_x] = \sum_{j=1}^{\infty} w_j(x)\tilde{\mu}_j(x). \tag{5.3}$$

We reiterate that instead of having a "simple" distribution about this mean, which is usually assumed to be normal, model (5.2) allows flexible error distributions.

As discussed in Chapter 2, the key differences distinguishing different proposals of form (5.2) present in literature are in the descriptions of the weight, mean, and variance functions. Most proposals assume a constant variance function, $\tilde{\sigma}_j^2(x) = \tilde{\sigma}_j^2$, with an additional simplified structure for the weights or mean functions. These simplifications are assumed because the model still remains highly flexible and maintains desirable properties such as large support and posterior consistency (MacEachern [2000], Barrientos et al. [2012], Pati et al. [2012], Norets and Pelenis [2012b]), yet computations and interpretations are much easier.

Models with constant weights, $w_j(x) = w_j$, and flexible mean functions were discussed in Section 2.3.3. The simplest proposal assumes a linear mean function, $\tilde{\mu}_j(x) = \underline{X}\tilde{\beta}_j$ with the prior specification of (w_j) defined by the Dirichlet process. We will denote this simple DP mixture model by DPM. References for the DPM include West et al. [1994], De Iorio et al. [2009], and Jara et al. [2010]. More flexible proposals extend this model by defining flexible mean functions, for example, through Gaussian processes (Gelfand et al. [2005]) or linear combinations of basis functions (De Iorio et al. [2004]), or by an alternative prior specification of the weights, for

example, through a two-parameter Poisson-Dirichlet process (Jara et al. [2010]). Clearly, computational complexity increases with a more flexible mean structure.

Instead, models with flexible weights and simple mean functions typically assume $\tilde{\mu}_j(x) = \underline{X}\tilde{\beta}_j$. A review of proposals for covariate-dependent weights is provided in Section 2.3.4, and references include include Griffin and Steele [2006], Dunson and Park [2008], Ren et al. [2011], and Rodriguez and Dunson [2011]. In addition, a novel proposal will be discussed in Chapter 6. Models based on the joint approach also imply flexible weights. These models were reviewed in Section 2.2 and further discussed in Chapter 4. They include a model also for x, which leads to some disadvantages; in particular, too much emphasis is placed on fitting the marginal of x. But, computations are much easier. The basic model based on the joint approach assumes the joint density of (Y, X) is a DP mixture (joint DPM).

Clearly, a crucial modeling aspect is the choice between constant and covariate-dependent weights. Thus, the first step of our study is a comparison between models with constant or covariate-dependent weight functions, when the focus is prediction, or estimation of the regression function. To simplify the analysis, we will assume x is continuous and univariate. We will compare the DPM, as the basic model of the form (5.2) with constant weight functions, and the joint DPM model, as the computationally simplest model with covariate-dependent weights.

The choice of the weight function is indeed crucial for the predictive performance of the model. The weight functions have implications on the latent partition of the data in different mixture components, and prediction is strongly dependent on such partition.

Models with constant weight functions implicitly assume that the covariates are not informative on the cluster allocation. This may be appropriate when the clustering is meant to model multiple response behavior that holds across the entire covariate space. However, when the real regression function cannot be captured by form specified by a single mean function, we show that (surprisingly) poor and uninformative prediction

may result. This occurs because in order to fit the data, the clusters will be associated to regions of the covariate space. The prediction is then a mixture of all the cluster-specific fitted regression curves, independent of x_{n+1} and the location of the clusters in the covariate space.

When the aim is estimation of the regression function, one imagines the clustering aims at selecting different curves, from the collection of available curves $\mu_j(\cdot)$, in different regions of the covariate space, for local approximation of the unknown regression curve. Models that allow for covariate-dependent weights encourage partitions which reflect this situation by implicitly using a notion of covariate-proximity clustering. In this case, the prediction is greatly improved; for a given partition, predictions based on clusters which are close to x_{n+1} in the covariate space have greater influence. The conditional predictions are then averaged across all partitions, according to the posterior distribution. Unfortunately, as we will illustrate, the information about what are reasonable, proximity-based partitions gets (dramatically) spread out in the posterior, leading to predictions based on undesirable partitions having too much impact and predictions based on desirable partitions with not enough impact.

These difficulties arise due to the huge number of partitions on which nonparametric mixture models assign a prior distribution. In particular, for both models choices, any partition of the n data points into k groups for $k = 1, \ldots, n$ is possible. There are

$$S_{n,k} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \begin{pmatrix} k \\ j \end{pmatrix} (k-j)^n,$$

a Stirling number of the second kind, ways to partition the n data points in to the k groups, and

$$B_n = \sum_{k=1}^{n} S_{n,k},$$

a Bell number, possible partitions of the n data points. Even for small n, this number is very large.

Many of these partitions are similar, differing only in a few subjects, and will provide a similar fit to the data. As a result, a large number of

partitions will adequately fit the data. The covariates, however, typically provide information on the partition structure and can be used to rule out some of these partitions. Our main point is that this information needs to be included in the prior probability law on the random partition, since it would otherwise be (dramatically) spread out in the posterior, due to the huge dimension of the partition space. In particular, if the aim is estimation of the regression function and the covariates are informative, partitions that satisfy an ordering constraint of the (x_i) are appropriate, as they strictly enforce the idea of covariate proximity and reflect the idea of clustering as tool for local approximation of the regression curve. Under this constraint, we can reduce the total number of partitions to just 2^{n-1} of the B_n total partitions. For example, for $n = 10$, the total number of partitions under this constraint is 512 of $115,975$ partitions, which is just 0.44% of the total partitions, and for $n = 100$ the percentage of partitions under this constraint is less than $10^{-83}\%$ of the total partitions. To not deal with this $10^{-83}\%$ would be unreasonable.

To resolve this issue, we propose to modify the distribution of the latent partition to rule out the undesirable partitions by setting the probability of these events to be zero, while still maintaining properties of the DP, such as the prior for k_n, the number of groups in a sample of size n. This allows the distribution of the partition to depend on the covariate according to the designated clustering principle and greatly reduces the number of possible partitions. Our aim is to demonstrate greatly improved prediction.

In general, ideas for reasonable configurations need to be given prominence, yet this can not be left to the chance of the route of any MCMC algorithm. But it is also very difficult to control the mass on the configurations in the prior to ensure there is sufficient mass on the desirable configurations in the posterior. It is only by putting zero mass on the undesirable configurations that we are able to ensure that there is appropriate posterior mass on the desirable configurations.

The research in this chapter is motivated by the problem of estimating the probability of Alzheimer's disease as a function of asymmetry of the

hippocampus. Nonparametric flexibility is needed to recover the non-monotone curve.

The chapter is organized as follows. In Section 5.2, we discuss predictive properties of the DPM and joint DPM models. In Section 5.3, we recalibrate the DPM to remove undesirable partitions and obtain useful posterior and predictive distributions. Section 5.4 covers the computational procedures for sampling and prediction under the modified DPM model. In Section 5.5, extensions to non-continuous and multivariate data are explored. Finally, numerical illustrations are presented in Section 5.6, and an application to predict AD status of subjects is presented in Section 5.7.

5.2 DPM and joint DPM models

5.2.1 DPM model

The DP mixture model for the distribution of response, Y_i, given the covariate, x_i, for $i = 1, \ldots, n$, has the form

$$\begin{aligned} Y_i|x_i, \beta_i, \sigma_i^2 &\overset{ind}{\sim} \mathrm{N}(\underline{X}_i\beta_i, \sigma_i^2), \\ (\beta_i, \sigma_i^2)|P &\overset{i.i.d.}{\sim} P, \\ \mathbf{P} &\sim \mathrm{DP}(\alpha P_0), \end{aligned} \tag{5.4}$$

Here, the base measure, P_0, is the conjugate multivariate normal–inverse gamma distribution, i.e $\beta|\sigma^2 \sim \mathrm{N}(\beta_0, \sigma^2 C^{-1})$ and $\sigma^2 \sim \mathrm{IG}(a, b)$, for some selection of (β_0, C, a, b).

The DPM model can be separated into a random partition model and a sampling model. Recall that $\rho_n = (s_1, \ldots, s_n)$ denotes the partition, where $s_i = j$ if (β_i, σ_i^2) is equal to the j^{th} unique parameter pair $(\beta_j^*, \sigma_j^{2*})$. The number of unique parameters is k, and n_j denotes number of parameters pairs that are equal to the j^{th} unique value. The random partition model

is obtained from the Pólya urn scheme;

$$p(\rho_n) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}\alpha^k \prod_{j=1}^{k} \Gamma(n_j).$$

The model is completed with the sampling model for the response given the partition and the covariate. From (5.4), we have independence across clusters and exchangeability within cluster, where within cluster a simple linear model is assumed.

Notice that the partition of the n observations is independent of x. This means that given the covariates, positive mass is assigned to any possible partition of the n observations into k groups and that *apriori* there is no preference for clusters with similar covariates.

The posterior of the partition given the observed data is proportional to the random partition model times the sampling model. The use of conjugate base measures in (5.4) allows for a closed form expression for the sampling model, and combining this expression with the prior, implies the posterior of partition is

$$p(\rho_n|y_{1:n}, x_{1:n}) \propto \alpha^k \prod_{j=1}^{k} \Gamma(n_j) \left(\frac{|C|}{|C + \underline{X}_j^{*\prime}\underline{X}_j^*|} \right)^{1/2} \frac{b^a \Gamma(a + n_j/2)}{\Gamma(a)(b + V_j^2/2)^{a+n_j/2}}, \tag{5.5}$$

where

$$V_j^2 = (y_j^* - \widehat{y}_j^*)' W_j (y_j^* - \widehat{y}_j^*),$$
$$W_j = (I_{n_j} - \underline{X}_j^*(C + \underline{X}_j^{*\prime}\underline{X}_j^*)^{-1}\underline{X}_j^{*\prime}),$$
$$\widehat{y}_j^* = \underline{X}_j^*\beta_0.$$

and y_j^* denotes the response of data points in cluster j, $\underline{X}_j^*$ is a matrix whose rows consist of $\underline{X}_i$ for data points in cluster j, and I_{n_j} denotes the n_j-dimensional identity matrix.

Equation (5.5) shows that *aposteriori* partitions with similar linear relationships between y and x are preferred.

Due to the large number of possible partitions, direct computation of (5.5) is unfeasible and requires MCMC approximations. We let $s =$

$1, \ldots, S$ index the iterations of a MCMC output, $\{\rho_n^s\}_{s=1}^S$, where for each s, ρ_n^s is an approximate sample from the posterior distribution of $[\rho_n|y_{1:n}, x_{1:n}]$. Due to the huge dimension of the partition space, many partitions will provide a good fit to the data, causing the chain to visit too many partitions with each one only visited very few times.

Under quadratic loss, the estimated regression curve at x_{n+1} corresponds to the point prediction of Y at x_{n+1}:

$$\widehat{m}(x_{n+1}) = \mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}].$$

Let $\mathcal{P}_n$ denote the set of all partitions of $\{1, \ldots, n\}$ and $\mathcal{P}(\rho_n) = \{1, ..., k+1\}$ denote the possible labels for the new data point given ρ_n; then, since *apriori* the random partition does not depend on the covariates,

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n \in \mathcal{P}_n} [\ldots] p(\rho_n|y_{1:n}, x_{1:n}), \tag{5.6}$$

$$[\ldots] = \sum_{s_{n+1} \in \mathcal{P}(\rho_n)} \mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}, \rho_{n+1}] p(s_{n+1}|\rho_n). \tag{5.7}$$

The inner term, (5.7), of (5.6), the prediction given ρ_n, is simply an average of all cluster-specific predictions which weights given by the Pólya urn scheme;

$$\mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}, \rho_n] = \frac{\alpha}{\alpha + n} \underline{X}_{n+1} \beta_0 + \sum_{j=1}^{k} \frac{n_j}{\alpha + n} \underline{X}_{n+1} \widehat{\beta}_j, \tag{5.8}$$

where

$$\widehat{\beta}_j = (C + \underline{X}_j^{*\prime} \underline{X}_j^*)^{-1} (C\beta_0 + \underline{X}_j^{*\prime} y_j^*)$$

is a vector containing the estimated intercept and slope for the regression line under the standard linear model given the response and covariates of subjects in cluster j.

Equation (5.8) shows that given the partition, the cluster-specific predictions are weighted according to the size of each cluster. This means that even if the new x_{n+1} is very far from the largest group, it is more likely to share the same regression line because many observations fall in that group. This aspect can clearly lead to very poor prediction.

Using equation (5.8), the expression for the regression curve estimate given in (5.6) becomes

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n \in \mathcal{P}_n} \left(\frac{\alpha}{\alpha+n} \underline{X}_{n+1}\beta_0 + \sum_{j=1}^{k} \frac{n_j}{\alpha+n} \underline{X}_{n+1}\widehat{\beta}_j \right) p(\rho_n | y_{1:n}, x_{1:n}),$$

which can be approximated through MCMC by

$$\widehat{m}(x_{n+1}) \approx \frac{1}{S} \sum_{s=1}^{S} \left(\frac{\alpha}{\alpha+n} \underline{X}_{n+1}\beta_0 + \sum_{j=1}^{k^s} \frac{n_j^s}{\alpha+n} \underline{X}_{n+1}\widehat{\beta}_j^s \right). \tag{5.9}$$

Thus, the prediction is averaged across all partitions, with weights given by their (estimated) posterior probability, and will therefore suffer from the issues for the posterior of the partition, namely, the insufficiently large posterior mass of desirable partitions that satisfy the notion of covariate proximity and insufficiently small posterior mass of undesirable partitions. If the prediction is based on an undesirable partition, the estimated regression line and/or weights within cluster be will be incorrect and the poor prediction resulting from this undesirable partition will be used in computations of (5.9). These issues are illustrated with examples in Section 5.6.

Also note that factoring out the $\underline{X}_{n+1}$ yields

$$\widehat{m}(x_{n+1}) = \underline{X}_{n+1} \left(\frac{\alpha}{\alpha+n}\beta_0 + \sum_{\rho_n \in \mathcal{P}_n} \sum_{j=1}^{k} p(\rho_n | y_{1:n}, x_{1:n}) \frac{n_j}{\alpha+n} \widehat{\beta}_j \right).$$

Thus, the curve estimate is merely a linear function of x_{n+1}, meaning that no matter where x_{n+1} lies in the covariate space, the same linear function is used to estimate y_{n+1}.

5.2.2 Joint DPM model

The joint DPM model was discussed in detail in Chapter 4, where the emphasis was on predictive properties of the model for an increasing number of covariates. Here we provide another detailed analysis of the joint DPM

model, but we focus on the case when the local model for Y is the standard linear regression model and examine the impact of the huge dimension of the partition space. The model is similar to (5.4), but also incorporates a model for the covariate,

$$\begin{aligned} Y_i|x_i, \beta_i, \sigma_i^2 &\overset{ind}{\sim} \mathrm{N}(\underline{X}_i\beta_i, \sigma_i^2), \\ X_i|\psi_i &\overset{ind}{\sim} F_x(\cdot|\psi_i), \\ (\beta_i, \sigma_i^2, \psi_i)|P &\overset{i.i.d.}{\sim} P, \\ \mathbf{P} &\sim \mathrm{DP}(\alpha P_{0Y} \times P_{0X}), \end{aligned}$$

where P_{0Y} is the base measure for the Y parameters and P_{0X} is the base measure for the X parameters. We assume the same structure for P_{0Y}, namely, the conjugate multivariate normal–inverse gamma for some selection of (β_0, C, a, b), and do not assume a specific form for P_{0X}, but for the examples in Section 5.6, where F_x is the normal distribution function, it is chosen to be the conjugate normal–inverse gamma.

Park and Dunson [2010] show that this model leads to the following covariate-dependent random partition model:

$$p(\rho_n|x_{1:n}) \propto \alpha^k \prod_{j=1}^{k} \Gamma(n_j) \int \prod_{\{i \in S_j\}} K(x_i; \psi) dP_{0X}(\psi), \tag{5.10}$$

where $S_j = \{i : s_i = j\}$ and $K(\cdot; \psi)$ is the density of F_x.

Müller and Quintana [2010] independently constructed a similar model, but were motivated by directly modifying the cohesion term of the random partition model by a factor that favors clusters with similar covariates. For the DPM model, the covariate-dependent random partition model is given by

$$p(\rho_n|x_{1:n}) \propto \alpha^k \prod_{j=1}^{k} \Gamma(n_j) g_x(x_j^*),$$

where $x_j^* = \{x_i\}_{i \in S_j}$.

The similarity function, $g_x(\cdot)$, captures the closeness of covariates, where large values indicate high similarity. Müller and Quintana [2010]

show that if the similarity function satisfies invariance with respect to permutations of the covariates and scalability, i.e

$$\int g_x(x_j^*, x)dx = g_x(x_j^*),$$

then

$$g_x(x_j^*) = \int \prod_{\{i \in S_j\}} K(x_i; \psi) dP_{0X}(\psi)$$

and the covariate-dependent random partition model is equivalent to that obtained in (5.10).

Even though (5.10) still assigns positive mass to any possible partition of the n subjects into k groups, clusters with similar covariates are encouraged. In particular, $K(\cdot; \psi)$ and P_{0X} together define a similarity function that measures the closeness of covariates, and multiplying by this function increases the probability of the desired clusters.

The posterior of the covariate-dependent partition is

$$p(\rho_n | y_{1:n}, x_{1:n}) \propto \alpha^k \prod_{j=1}^{k} \Gamma(n_j) g_x(x_j^*) \left(\frac{|C|}{|C + \underline{X}_j^{*\prime} \underline{X}_j^*|} \right)^{1/2} * \frac{b^a \Gamma(a + n_j/2)}{\Gamma(a)(b + V_j^2/2)^{a+n_j/2}}.$$

Due to incorporation of the similarity function, desirable partitions that satisfy the notion of covariate proximity will have higher posterior mass, undesirable partitions will have smaller posterior mass, and the MCMC chain will visit more reasonable partitions. However, the total number of partitions has not changed; undesirable partitions still have positive prior mass, and incorporation of the similarity function may not be enough to ensure their posterior mass is sufficiently small. In particular, many of the undesirable partitions will differ from a desirable partition in only a few subjects and may, thus, fit the data adequately, even though knowledge of the covariates implies superiority of the desirable partition, particularly in terms of improved prediction. This may not only cause the posterior mass of such undesirable partitions to be too large but will also result in a diluted posterior mass of the desirable partitions.

For this model, since the random partition depends on the covariates, the expression used to compute the prediction of y_{n+1} given x_{n+1} and the data, is slightly different than that used for the DPM model (equation (5.6));

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n \in \mathcal{P}_n} [\ldots] p(\rho_n | y_{1:n}, x_{1:n}),, \tag{5.11}$$

$$[\ldots] = \sum_{s_{n+1} \in \mathcal{P}(\rho_n)} \mathrm{E}[Y_{n+1} | y_{1:n}, x_{1:n+1}, \rho_{n+1}] p(s_{n+1} | \rho_n, x_{1:n+1}) c_0 \tag{5.12}$$

where $c_0 = f(x_{n+1}|\rho_n, y_{1:n}, x_{1:n})/f(x_{n+1}|y_{1:n}, x_{1:n})$. The term c_0 needs to be included because $p(\rho_n \mid y_{1:n}, x_{1:n+1})$ is no longer equal to $p(\rho_n \mid y_{1:n}, x_{1:n})$. Furthermore, notice that the predictive distribution of s_{n+1} now depends on $x_{1:n}$ and x_{n+1}.

The inner term, (5.12), of (5.11) is again an average of all cluster-specific predictions but the weights given by the Pólya urn scheme are modified by the cluster-specific predictive densities of x_{n+1};

$$\frac{\alpha}{c_1} g_x(x_{n+1}) \underline{X}_{n+1} \beta_0 + \sum_{j=1}^{k} \frac{n_j}{c_1} g_x(x_{n+1}|x_j^*) \underline{X}_{n+1} \widehat{\beta}_j, \tag{5.13}$$

where $c_1 = f(x_{n+1}|y_{1:n}, x_{1:n})/(\alpha + n)$ and

$$g_x(x_{n+1}|x_j^*) = \int K(x_{n+1}; \psi) dP_{0X}(\psi|x_j^*)$$

is the predictive density of x_{n+1} given the x-observations in the j^{th} cluster.

The cluster-specific predictive density of x_{n+1} measures the closeness of x_{n+1} and the clusters in the covariate space. From expression (5.13), we see that regression lines for clusters close to x_{n+1} in covariate space are assigned more weight. However, regression lines for clusters far from x_{n+1} in the covariate space still have positive weight resulting unnecessary inclusion of poor predictions based on these clusters in the average computed in (5.13).

The final expression for the prediction of y_{n+1} given x_{n+1} and the data, i.e. the regression curve estimate, is given by

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n \in \mathcal{P}_n} [\ldots] p(\rho_n | y_{1:n}, x_{1:n}),$$

$$[\ldots] = \frac{\alpha}{c_1} g_x(x_{n+1}) \underline{X}_{n+1} \beta_0 + \sum_{j=1}^{k} \frac{n_j}{c_1} g_x(x_{n+1} | x_j^*) \underline{X}_{n+1} \widehat{\beta}_j,$$

which is approximated by

$$\widehat{m}(x_{n+1}) \approx \frac{1}{S} \sum_{s=1}^{S} \left(\frac{\alpha}{\hat{c}_1} g_x(x_{n+1}) \underline{X}_{n+1} \beta_0 + \sum_{j=1}^{k^s} \frac{n_j^s}{\hat{c}_1} g_x(x_{n+1} | x_j^{*s}) \underline{X}_{n+1} \widehat{\beta}_j^s \right), \tag{5.14}$$

where

$$\hat{c}_1 = \frac{1}{S} \sum_{s=1}^{S} \alpha g_x(x_{n+1}) + \sum_{j=1}^{k^s} n_j^s g_x(x_{n+1} | x_j^{*s}).$$

Again, the estimate obtained in (5.14) by averaging over all partitions visited by the chain will suffer from the issues for the posterior of the partition mentioned above as well as poor prediction arising from undesirable partitions with insufficiently small posterior mass.

Finally, note that the regression curve estimate is no longer a linear function of x_{n+1}, since the weights assigned to each regression line depend on x_{n+1}.

5.3 A restricted DPM model

In regression settings when the aim is estimation of the regression function and the covariates are informative for prediction, partitioning should be based on the proximity of the covariates to reflect the notion of local approximation of the regression curve. Due to the unrestricted nature of the clusters offered by nonparametric mixture models, this idea of covariate proximity needs to be specifically enforced on the partition structure.

When the covariate is univariate, the idea of covariate proximity is naturally expressed by the ordering of x. For example, if $x_i < x_{i'} < x_{i''}$, it is reasonable to assume that if subjects (i, i'') are clustered together, then subject i' is also in that cluster. To this aim, we use the natural ordering of x to determine the allowed partitions and remove undesirable partitions by adjusting the conditional density of partition given the covariate, so that their mass is zero.

Let $x_{(1)}, \ldots, x_{(n)}$ denote the ordered values of $x_1, \ldots, x_n$, and $y_{(1)}, \ldots, y_{(n)}$ and $s_{(1)}, \ldots, s_{(n)}$ be the corresponding values of $y_1, \ldots, y_n$ and $s_1, \ldots, s_n$. The distribution of the partition implied by the DP is invariant to a relabelling of the clusters as long as the partition is preserved. This means that we can relabel the clusters, so that the subject with the smallest covariate is in the first cluster. To impose the order constraint that if subjects i and i'' are clustered together, then all subjects whose covariates are between x_i and $x_{i''}$ are in the same cluster, we require that

$$s_{(1)} \leq \ldots \leq s_{(n)}. \tag{5.15}$$

Unfortunately, while simply multiplying $p(\rho_n|x_{1:n})$ by the indicator that $s_{(1)} \leq \ldots \leq s_{(n)}$, an approach similar to the one used in Fuentes-Garcia et al. [2010], does remove the unwanted partitions, it also leads to an undesirable prior for k. Such an approach would cause the prior for k to place a high mass on $k = 1$, and for a fixed value of α, the mass assigned to $k = 1$ increases with the sample size. This strange effect is due to the fact that we are removing no partitions for $k = 1$ and $k = n$ and many as $k \to n/2$. The mass of the removed partitions is spread out evenly among the remaining partitions, thus increasing the relative weight of $k = 1$ and $k = n$ and decreasing the relative weight of moderate values of k.

To avoid this effect, we define a covariate-dependent random partition model that both removes undesirable partitions and retains the DP's prior for k, as is demonstrated in the following proposition.

Proposition 5.3.1 *The probability measure on the random partition de-*

fined by

$$p^*(\rho_n|x_{1:n}) = \frac{\Gamma(\alpha)\Gamma(n+1)}{\Gamma(\alpha+n)}\frac{\alpha^k}{k!}\prod_{j=1}^{k}\frac{1}{n_j} * \mathbf{1}(s_{(1)} \leq \ldots \leq s_{(n)}) \qquad (5.16)$$

satisfies the order constraint (5.15) and has the same marginal for k, as that induced by the Dirichlet process.

Proof. Clearly, by construction, the random partition model satisfies the order constraint (5.15). Thus, we only need to prove that the marginal for k is equivalent to that induced by the DP. The proof relies on the fact that under constraint (5.15), the partition is uniquely determined by $(n_1, \ldots, n_k, k)$. In particular, if $s_{(1)} \leq \ldots \leq s_{(n)}$ and $\mathbf{n_1} = n_1, \ldots \mathbf{n}_k = n_k, \mathbf{k} = k$, then

$$\mathbf{s}_{(1)} = 1, \ldots, \mathbf{s}_{(n_1^+)} = 1, \ldots, \mathbf{s}_{(n_{k-1}^+ + 1)} = k, \ldots \mathbf{s}_{(n_k^+)} = k.$$

Alternatively, if $s_{(1)} \leq \ldots \leq s_{(n)}$ and $\mathbf{s}_{(1)} = s_{(1)}, \ldots, \mathbf{s}_{(n)} = s_{(n)}$, then

$$\mathbf{n}_1 = \sum_{i=1}^{n}\mathbf{1}(s_{(i)} = 1), \ldots, \mathbf{n}_k = \sum_{i=1}^{n}\mathbf{1}(s_{(i)} = k), \mathbf{k} = 1 + \sum_{i=1}^{n-1}\mathbf{1}(s_{(i)} < s_{(i+1)}).$$

This implies that

$$p^*(n_1, \ldots, n_k, k|x_{1:n}) = \frac{\Gamma(\alpha)\Gamma(n+1)}{\Gamma(\alpha+n)}\frac{\alpha^k}{k!}\prod_{j=1}^{k}\frac{1}{n_j}. \qquad (5.17)$$

The prior for $\{m_i\}$, the number of clusters of size i for $i = 1, ..., n$, can be obtained by summing (5.17) over the set of $(n_1, \ldots, n_k)$ that satisfy $m_1, \ldots, m_n$. This set is given by $(n_{\pi(1)}, \ldots, n_{\pi(k)})$ for any permutation π of the cluster indices, where $(n_1, \ldots, n_k)$ is a specific vector that satisfies $m_1, \ldots, m_n$. Since (5.17) is invariant to a permutation of cluster indices, the probability of $m_1, \ldots, m_n$ is simply the probability of a specific $(n_1, \ldots, n_k)$ that satisfies $m_1, \ldots, m_n$ multiplied by the number of unique ways to order the m_i clusters of size i for $i = 1, \ldots, n$, which is

$$\frac{k!}{\prod_{i=1}^{n} m_i!}.$$

This implies that

$$p^*(m_1,\ldots,m_n|x_{1:m}) = \frac{\Gamma(\alpha)\Gamma(n+1)}{\Gamma(\alpha+n)}\frac{k!}{\prod_{i=1}^n m_i!}\frac{\alpha^k}{k!}\prod_{j=1}^k\frac{1}{n_j}$$
$$= \frac{\Gamma(\alpha)\Gamma(n+1)}{\Gamma(\alpha+n)}\alpha^k\frac{1}{\prod_{i=1}^n i^{m_i}m_i!}.$$

This is prior for $\{m_i\}$ induced by the DP (see Antoniak [1974]). Since, $k=\sum_{i=1}^n m_i$, it follows that prior for k is equivalent to that of the DP. ∎

Notice that the proof of this proposition shows that the random partition model (5.16) has a stronger resemblance to random partition model of the DP; it maintains the prior for $\{m_i\}$, the number of clusters of size i for $i=1,...,n$. Then, since $k=\sum_i^n m_i$, the equivalence of the prior of k follows. The proof relies on the fact that under constraint (5.15), the partition is uniquely determined by $(n_1,\ldots,n_k,k)$, a property that will also be exploited for computations.

This simple construction only allows for clusters with similar x, greatly reduces the total number of partitions, and ensures undesirable partitions have zero posterior mass. We note that this model can recover a wide regression functions, including functions which discontinuities or sharp changes. However, as the number discontinuities increases or changes in the function become more rapid, we expect more data points will be required for a good estimation.

5.3.1 The posterior distribution

The posterior distribution of the partition is

$$p^*(\rho_n|y_{1:n},x_{1:n}) \propto \frac{\alpha^k}{k!}\prod_{j=1}^k\frac{1}{n_j}\left(\frac{|C|}{|C+\underline{X}_j^{*\prime}\underline{X}_j^*|}\right)^{1/2} * \frac{b^a\Gamma(a+n_j/2)}{\Gamma(a)(b+V_j^2/2)^{a+n_j/2}}$$
$$* \mathbf{1}(s_{(1)}\leq\ldots\leq s_{(n)}),$$

which depends on the hyper-parameters; (α,C,b,a). The interpretation of these parameters is similar to the DP model. A large value for α will

encourage more clusters through the factor of α^k. For a given k, the term $\prod_{j=1}^{k} n_j^{-1}$ will favor partitions with one large cluster and several small clusters. Thus, if one believes that *apriori* the clusters are balanced, the prior distribution of the partition should be adjusted appropriately.

Given σ^2, the prior variance–covariance matrix of the intercept and slope is $\sigma^2 C^{-1}$. Typically, C is a diagonal matrix with small values on the diagonal so that the prior is non-informative. In this case, $|C| < 1$ and

$$\prod_{j=1}^{k} \left(\frac{|C|}{|C + \underline{X}_j^{*\prime} \underline{X}_j^*|} \right)^{1/2} \approx \frac{|C|^{k/2}}{\prod_{j=1}^{k} |\underline{X}_j^{*\prime} \underline{X}_j^*|^{1/2}}.$$

The term $|C|^{k/2}$ will discourage a large number of clusters, while

$$\prod_{j=1}^{k} |\underline{X}_j^{*\prime} \underline{X}_j^*|^{1/2} = \prod_{j=1}^{k} n_j \left(\sum_{i \in S_j} \frac{(x_i - \bar{x}_j)^2}{n_j} \right)^{1/2},$$

where $\bar{x}_j$ is the sample mean of the (x_i) in cluster j, will encourage clusters with similar values of the covariate and unbalanced clusters. For a given k, the term $\prod_{j=1}^{k} \Gamma(a + n_j/2)/\Gamma(a)$ will also encourage unbalanced clusters. Finally, $\prod_{j=1}^{k} b^a/(b + V_j^2/2)^{a+n_j/2}$ will encourage clusters with similar values of the covariate and similar linear response curve, since V_j^2 will be smaller in this case.

5.3.2 Prediction

Given the partition of the observed subjects and new subject, the predictive distribution has a known form and can be easily computed and sampled from. In particular, suppose that according to ρ_{n+1} the new subject is in cluster j. Then, the predictive distribution of Y_{n+1} is obtained from standard computations based on the observations in cluster j. In particular, it is a non-central t-distribution with location $\underline{X}_{n+1} \widehat{\beta}_j$, scale $\widehat{b}_j^{-1} \widehat{a}_j W_{n+1,j}$, and $2a + n_j$ degrees of freedom:

$$(Y_{n+1} - \underline{X}_{n+1} \widehat{\beta}_j) * \left(\frac{\widehat{a}_j W_{n+1,j}}{\widehat{b}_j} \right)^{1/2} \mid \rho_{n+1}, y_{1:n}, x_{1:n} \sim \mathcal{T}(2a + n_j),$$

where $\mathcal{T}(\nu)$ denotes the t-distribution with ν degrees of freedom. Here we denote the number of observed subjects in cluster j by n_j and the response and covariate matrix for the n_j observed subjects in cluster j by $(\underline{X}_j^*, y_j^*)$, we define

$$W_{n+1,j} = 1 - \underline{X}_{n+1}(\widehat{C}_j + \underline{X}'_{n+1}\underline{X}_{n+1})^{-1}\underline{X}_{n+1},$$

$$\widehat{C}_j = C + \underline{X}_j^{*\prime}\underline{X}_j^*,$$

$$\widehat{a}_j = a + n_j/2, \text{ and } \widehat{b}_j = b + V_j^2/2,$$

and compute $\widehat{\beta}_j$ and V_j^2 based on $(y_j^*, \underline{X}_j^*)$. If the new subject belongs to a new cluster, then $n_j = 0$ and the updated parameters, $\widehat{a}_j, \widehat{b}_j, \widehat{\beta}_j, \widehat{C}_j$ are given by the prior parameters.

Define $\mathcal{C}_n$ as the set of possible partitions of the n subjects under the restricted DPM model and $\mathcal{C}(\rho_n)$ as the set of values for s_{n+1} such that ρ_{n+1} restricted to n observed subjects is ρ_n. The predictive mean of Y_{n+1} is given in the following proposition.

Proposition 5.3.2 *If the random partition model is described by (5.16), then the prediction of y_{n+1} given x_{n+1} and the data is*

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n \in \mathcal{C}_n} [\ldots] p^*(\rho_n \mid y_{1:n}, x_{1:n}), \tag{5.18}$$

$$[\ldots] = \sum_{s_{n+1} \in \mathcal{C}(\rho_n)} E[Y_{n+1} \mid y_{1:n}, x_{1:n+1}, \rho_{n+1}] \frac{p^*(\rho_{n+1} \mid x_{1:n+1})}{p^*(\rho_n \mid x_{1:n})} c_2, \tag{5.19}$$

where the inner term, (5.19), of (5.18) is

$$= \begin{cases} \frac{\alpha}{c_3(k+1)}\underline{X}_{n+1}\beta_0 + \frac{n_j}{c_3(n_j+1)}\underline{X}_{n+1}\widehat{\beta}_j & \text{if } x_{n+1} < x_{(1)} \text{ or } x_{n+1} > x_{(n)}, \\ \frac{\alpha}{c_3(k+1)}\underline{X}_{n+1}\beta_0 + \frac{n_j}{c_3(n_j+1)}\underline{X}_{n+1}\widehat{\beta}_j & \text{if } x_{(i)} < x_{n+1} < x_{(i+1)} \text{ and} \\ \quad + \frac{n_{j+1}}{c_3(n_{j+1}+1)}\underline{X}_{n+1}\widehat{\beta}_{j+1} & s_{(i)} = j, s_{(i+1)} = j+1, \\ \frac{n_j}{c_3(n_j+1)}\underline{X}_{n+1}\widehat{\beta}_j & \text{if } x_{(i)} < x_{n+1} < x_{(i+1)} \text{ and} \\ & s_{(i)} = j, s_{(i+1)} = j, \end{cases}$$

with $c_2 = p^*(y_{1:n} \mid x_{1:n})/p^*(y_{1:n}|x_{1:n+1})$ and $c_3 = (\alpha + n)/(c_2(n+1))$.

Proof. The proof relies on simple computations. The prediction of the y_{n+1} is

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n \in \mathcal{C}_n} \sum_{s_{n+1} \in \mathcal{C}(\rho_n)} \mathrm{E}[Y_{n+1} \mid y_{1:n}, x_{1:n+1}, \rho_{n+1}] \\ * p^*(s_{n+1} \mid \rho_n, y_{1:n}, x_{1:n+1}) p^*(\rho_n \mid y_{1:n}, x_{1:n+1}).$$

The posterior of $[\rho_n \mid y_{1:n}, x_{1:n+1}]$ can be written in terms of the posterior of $[\rho_n \mid y_{1:n}, x_{1:n}]$, since

$$\begin{aligned} p^*(\rho_n \mid y_{1:n}, x_{1:n+1}) &= \frac{p^*(\rho_n \mid x_{1:n+1})}{p^*(\rho_n \mid x_{1:n})} \frac{p^*(\rho_n \mid x_{1:n})}{p^*(y_{1:n} | x_{1:n+1})} p(y_{1:n} \mid \rho_n, x_{1:n}) \\ &= \frac{p^*(\rho_n \mid x_{1:n+1})}{p^*(\rho_n \mid x_{1:n})} \frac{p^*(y_{1:n} \mid x_{1:n})}{p^*(y_{1:n} | x_{1:n+1})} p^*(\rho_n \mid y_{1:n}, x_{1:n}). \end{aligned}$$

Using a similar trick, the predictive density of $[s_{n+1} \mid \rho_n, y_{1:n}, x_{1:n+1}]$ can be written as

$$p^*(s_{n+1} \mid \rho_n, y_{1:n}, x_{1:n+1}) = \frac{p^*(\rho_{n+1} \mid x_{1:n+1})}{p^*(\rho_n \mid x_{1:n+1})}.$$

Combining these results leads to equation (5.18).

To compute (5.19), we need to consider the following three cases:

1. If x_{n+1} is an end point (i.e. $x_{n+1} < x_{(1)}$ or $x_{n+1} > x_{(n)}$), the ordering constraint implies that there are two possible partitions of the $n+1$ subjects whose restriction to the n observed subjects is ρ_n. Suppose $x_{n+1} < x_{(1)}$, then either (i) the new subject is in the first cluster with weight proportional to $\frac{n_1}{n_1+1}$, or (ii) the new subject is in a new cluster with weight proportional to $\frac{\alpha}{k+1}$.

2. If x_{n+1} lies between two subjects in different clusters, say clusters j and $j+1$, the ordering constraint implies that there are three possible partitions of the $n+1$ subjects whose restriction to the n observed subjects is ρ_n. Either (i) the new subject is in the cluster j with weight proportional to $\frac{n_j}{n_j+1}$, (ii) the new subject is in the cluster $j+1$ with weight proportional to $\frac{n_{j+1}}{n_{j+1}+1}$, or (iii) the new subject is in a new cluster with weight proportional to $\frac{\alpha}{k+1}$.

3. Otherwise, x_{n+1} lies between two subjects who are in the same cluster, and the ordering constraint implies that there is only one possible partition of the $n+1$ subjects whose restriction to the n observed subjects is ρ_n. The new subject is in the cluster j with weight proportional to $\frac{n_j}{n_j+1}$. ∎

Notice that the expression used to compute the prediction is slightly different than that used for the joint DPM model. This is because we do not require X to be stochastic, and therefore, we do not have a model for X in computation of the prediction. As for the joint DPM model, $p^*(\rho_n \mid y_{1:n}, x_{1:n+1}) \neq p^*(\rho_n \mid y_{1:n}, x_{1:n})$.

From Proposition 5.3.2, we see that given the partition, the prediction is an average of predictions based only on clusters close to x_{n+1} in the covariate space, where higher weight is given to neighbouring clusters with many individuals. Also, smaller α and larger k will give less weight to the prediction for a new cluster.

5.4 Computations

By enforcing an ordering constraint on the partition based on the covariate, we have reduced the number of possible partitions of n subjects into k groups from $S_{n,k}$, a Stirling number of the second kind, to $\begin{pmatrix} n-1 \\ k-1 \end{pmatrix}$; the first cluster must start with the first subject, and there are $\begin{pmatrix} n-1 \\ k-1 \end{pmatrix}$ ways to choose where to start following $k-1$ clusters among $n-1$ remaining subjects. Thus, the constraint imposed reduces the total number of partitions from B_n to

$$\sum_{k=1}^{n} \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} = 2^{n-1}.$$

However, for moderate to large n, this number is still large, and one needs to resort to MCMC methods to approximate $p^*(\rho_n | y_{1:n}, x_{1:n})$. To explore the space of partitions, we use the reversible jump MCMC Algorithm

as described in Fuentes-Garcia et al. [2010] and briefly described in the following paragraph.

First, recall that ρ_n is uniquely determined by $(n_1, \ldots, n_k, k)$. At each iteration, one of two types of moves is proposed: a split, where a group of size bigger than one is divided into two, so that k is increased by 1, or a merge, where two neighbouring groups are combined, so that k is decreased by 1. Uniform distribution are used for both types of moves, thus

$$p^*(n_1, \ldots, n_{k+1}, k+1 | n_1, \ldots, n_k, k) = \frac{1}{k_g(n_h - 1)},$$
$$p^*(n_1, \ldots, n_{k-1}, k-1 | n_1, \ldots, n_k, k) = \frac{1}{k-1},$$

where for a split, h is the group selected to split and k_g is the number of groups of size larger than one. Letting $n^{(k)} = (n_1, \ldots, n_k, k)$, the acceptance probabilities for a split or merge, respectively, are

$$a(n^{(k+1)} | n^{(k)}) = \min\left\{1, \frac{p^*(n^{(k+1)} | y_{1:n}, x_{1:n})}{p^*(n^{(k)} | y_{1:n}, x_{1:n})} \frac{k_g(n_h - 1)}{k}\right\},$$
$$a(n^{(k-1)} | n^{(k)}) = \min\left\{1, \frac{p^*(n^{(k-1)} | y_{1:n}, x_{1:n})}{p^*(n^{(k)} | y_{1:n}, x_{1:n})} \frac{k-1}{(k-1)_g(n_{h_1} + n_{h_2} - 1)}\right\},$$

where for a merge, (h_1, h_2) are the two groups selected to merge and $(k-1)_g$ is the number of groups of size larger than one under the proposed merged partition. The proposed move is then accepted with its corresponding acceptance probability. Next, a shuffle of the current partition is performed, where two adjacent groups of size (n_{h_1}, n_{h_2}) are merged and then split into two groups of size $(n^*_{h_1}, n^*_{h_2})$. The shuffle is accepted with probability

$$a(n^{(k)*} | n^{(k)}) = \min\left\{1, \frac{p^*(n^{(k)*} | y_{1:n}, x_{1:n})}{p^*(n^{(k)} | y_{1:n}, x_{1:n})}\right\}.$$

For prediction, we use the estimate of $p(\rho_n \mid y_{1:n}, x_{1:n})$ from the MCMC algorithm. We consider all (ρ_{n+1}) whose restriction to the observed n subjects is in the set of (ρ_n) with positive estimated posterior probabilities.

For each ρ_n^s visited in the chain, the local prediction $\underline{X}_{n+1}\widehat{\beta}_j^s$ and the non-normalized weight, denoted by $w_j^s(x_{n+1})$, are computed for $j \in \mathcal{C}(\rho_n^s)$. The prediction of y_{n+1} given x_{n+1} and the data, equation (5.18), can be estimated by

$$\widehat{m}(x_{n+1}) \approx \sum_{s=1}^{S} \sum_{j \in \mathcal{C}(\rho_n^s)} \frac{w_j^s(x_{n+1})}{\widehat{c}_3} \underline{X}_{n+1}\widehat{\beta}_j^s,$$

where

$$\widehat{c}_3 = \sum_{s=1}^{S} \sum_{j \in \mathcal{C}(\rho_n^s)} w_j^s(x_{n+1}).$$

Note that because we have greatly reduced the parameter space, we are able to sample the partition jointly as opposed to the DPM and joint DPM models which require sampling from the full conditional of cluster label for each subject. This results in much faster MCMC computations and better mixing.

5.5 Extensions

To illustrate our point, we have focused on regression with univariate and continuous data, but our discussion can be extended to more general regression problems. We show how to extend the proposed method to univariate regression with non-continuous data. As is common to many methods, such as splines, extensions for multivariate covariates are more complicated, but we outline the basic structure that would be required.

5.5.1 Extensions to non-continuous covariates

When subjects may have equal values of the covariate, a strict ordering of the covariates is no longer available, but, in most cases, a strict ordering of the unique values of the covariates is available. In particular, when the covariate is binary, ordinal, counts, or continuous with possible repeated values of the observed covariates (for example, due to rounding errors or

experiment design), an ordering of the unique covariate values is sensible. We demonstrate how to handle these cases.

Let k_x denote the number of unique values among the observed covariates, let $n_{x,h}$ denote the number of subjects with the h^{th} unique ordered covariate, for $h = 1, \ldots, k_x$, and let $\underline{s}_{(h)}$ denote a vector containing the labels for the $n_{x,h}$ subjects with the h^{th} unique ordered covariate.

In this setting, undesirable partitions are those which violate the constraint

$$\underline{s}_{(1)} \leq \ldots \leq \underline{s}_{(k_x)}, \tag{5.20}$$

where $\underline{s}_{(h)} \leq \underline{s}_{(h')}$ if $s_{(h,i)} \leq s_{(h',j)}$, for $i = 1, \ldots, n_{x,h}$ and $j = 1, \ldots, n_{x,h'}$. Again, we want to define a random partition model which both removes partitions violating (5.20) and maintains the DP's prior for k.

For the following proposition, we define $n^+_{x,h} = \sum_{l=1}^{h} n_{x,l}$ and $n^+_{x,0} = 0$, and similarly, $n^+_j = \sum_{l=1}^{j} n_l$ and $n^+_0 = 0$. Let

$$k_{x,h} = \sum_{j=1}^{k} \mathbf{1}(n^+_{x,h-1} \leq n^+_{j-1}) * \mathbf{1}(n^+_j \leq n^+_{x,h})$$

for $h = 1, \ldots, k_x$, denote the number of clusters which both start and end among subjects with the h^{th} unique ordered covariate.

Proposition 5.5.1 *The probability measure on the random partition defined by*

$$p^*(\rho_n | x_{1:n}) = \frac{\Gamma(\alpha)\Gamma(n+1)}{\Gamma(\alpha+n)} \frac{\alpha^k}{k!} \prod_{j=1}^{k} \frac{1}{n_j} * \prod_{h=1}^{k_x} k_{x,h}! * \mathbf{1}(\underline{s}_{(1)} \leq \ldots \leq \underline{s}_{(k_x)})$$
$$* \prod_{h=1}^{k_x} \prod_{j=1}^{k} \left(\frac{(n^+_j - \max(n^+_{x,h-1}, n^+_{j-1}))!(n^+_{x,h} - n^+_j)!}{(n^+_{x,h} - \max(n^+_{x,h-1}, n^+_{j-1}))!} \right)^{\mathbf{1}(n^+_{x,h-1} < n^+_j < n^+_{x,h})} \tag{5.21}$$

satisfies the order constraint (5.20) and has the same marginal for k, as that induced by the Dirichlet process.

Proof For $j = 1, ...k$, if n_j specifies a split within subjects with the h^{th} unique ordered covariate, define $S_{j,x}$ as the set of indices of subjects among those with the h^{th} unique ordered covariate in group j, i.e

$$S_{j,x} = \{i : s_{(h,i)} = j, n^+_{x,h-1} < n^+_j < n^+_{x,h}\}.$$

The set $S_{j,x}$ may be empty if $n^+_j = n^+_{x,h}$ for some h. If multiple clusters start and end among subjects with the same covariate, the clusters are ordered according to subject indices. Under the order constraint (5.20), it is straightforward to show that the partition is uniquely determined by $(n_1, \ldots, n_k, k)$ and the sets $S_{j,x}$. This implies that

$$p^*(n_1, S_{1,x}, \ldots, n_k, S_{k,x}, k|x_{1:n}) = \frac{\Gamma(\alpha)\Gamma(n+1)}{\Gamma(\alpha+n)} \frac{\alpha^k}{k!} \prod_{j=1}^{k} \frac{1}{n_j} * \prod_{h=1}^{k_x} k_{x,h}! \\ * \prod_{h=1}^{k_x} \prod_{j=1}^{k} \left(\frac{(n^+_j - \max(n^+_{x,h-1}, n^+_{j-1}))!(n^+_{x,h} - n^+_j)!}{(n^+_{x,h} - \max(n^+_{x,h-1}, n^+_{j-1}))!} \right)^{\mathbf{1}(n^+_{x,h-1} < n^+_j < n^+_{x,h})}. \tag{5.22}$$

Since (5.22) doesn't depend on $(S_{1,x}, \ldots S_{k,x})$, the marginal for $(n_1, \ldots, n_k, k)$ is obtained by multiplying (5.22) by the cardinality of $(S_{1,x}, \ldots S_{k,x})$. For $j = 1, ..., k$ such that $n^+_{x,h-1} < n^+_j < n^+_{x,h}$ for some h, there are

$$\begin{pmatrix} n^+_{x,h} - \max(n^+_{x,h-1}, n^+_{j-1}) \\ n^+_j - \max(n^+_{x,h-1}, n^+_{j-1}) \end{pmatrix}$$

ways to choose the $n^+_j - \max(n^+_{x,h-1}, n^+_{j-1})$ subjects with the h^{th} unique ordered covariate for group j. The cardinality is then given by the product of this number over j divided by $\prod_{h=1}^{k_x} k_{x,h}!$. This division is needed because simply taking the product does not account for ordering of clusters according to subject indices for $k_{x,h}$ clusters that both start and end among subjects with the h^{th} unique covariate. Thus,

$$p^*(n_1, \ldots, n_k, k|x) = \frac{\Gamma(\alpha)\Gamma(n+1)}{\Gamma(\alpha+n)} \frac{\alpha^k}{k!} \prod_{j=1}^{k} \frac{1}{n_j},$$

and the same arguments used in the proof of Proposition (5.3.1) can be applied to show the marginal prior for k is equivalent to that of the DP. ∎

Since, the partition is no longer completely determined by $(n_1, \ldots, n_k, k)$, the MCMC scheme needs to be modified appropriately to handle this.

Proposition 5.5.2 *The random partition model of the Dirichlet process is a special case of the covariate dependent random partition model defined by (5.21) when all covariates are equal.*

The proof of this proposition is straightforward. If all covariates are equal then $k_x = 1$, $k_{x,h} = k$, and $n^+_{x,1} = n$. After plugging in these values and noticing that

$$\prod_{j=1}^{k-1} \frac{n_j!(n - n_j^+)!}{(n - n_{j-1}^+)!} = \frac{1}{n!} \prod_{j=1}^{k} n_j!,$$

(5.21) reduces to the random partition model of the DP.

The nice property given in Proposition 5.5.2 is not satisfied by the joint DPM model. In fact, Müller and Quintana [2010] mention this as one of the undesirable features of the model.

A second approach to handle non-continuous covariates is to impose a further constraint requiring that the partition must also be ordered according to the response. Let

$$s_{(1,1)}, \ldots, s_{(1,n_{x,1})}, \ldots, s_{(k_x,1)}, \ldots, s_{(k_x,n_{x,k_x})}$$

denote the partition ordered first according to the covariate and then according the response.

In this case, one can use the covariate dependent random partition model of (5.21) with a slightly different sampling model,

$$f(y_{1:n}|\rho_n, x_{1:n}) \propto \prod_{j=1}^{k} f(y_{j,1}, \ldots, y_{j,n_j}|x_{j,1}, \ldots, x_{j,n_j}) \\ * \prod_{h=1}^{k_x} \mathbf{1}(s_{(h,1)} \leq \ldots \leq s_{(h,n_{x,h})}). \tag{5.23}$$

Since *aposteriori* the partition is now uniquely determined by the values of $(n_1, \ldots, n_k, k)$, and the MCMC algorithm discussed in Section 5.4 can be used to obtain posterior samples of the partition. However, for prediction, the sampling model is modified.

Following from Proposition 5.5.2, if the covariate dependent random partition model is defined by (5.21) and the sampling model is given by (5.23), then when all covariates are equal, the model reduces to a model similar to that of Fuentes-Garcia et al. [2010].

Both of the proposed methods for non-continuous covariates, are equivalent to the model in Section 5.3 when all covariates are distinct. We recommend use of the second method because a more imposing ordering constraint is used, resulting in a reduced number of possible partitions and a more identifiable model.

5.5.2 Extensions to non-continuous responses

If a closed form expression is available for the sampling model, extensions for a non-continuous response are straightforward. Once the expression for the sampling model is obtained, the MCMC algorithm in Section 5.4 can be used. When no closed form expression is available, extensions for a non-continuous response become more complicated.

Here, we demonstrate how to handle a binary response by building on local probit models. This model will be used in Section 5.7 to predict Alzheimer's disease status based on asymmetry of the hippocampus. Suppose the response for subject i, Y_i, is the indicator that the latent variable, $\tilde{Y}_i$, is positive, i.e. $Y_i = \mathbf{1}(\tilde{Y}_i > 0)$. The model for the latent $\tilde{Y}_i$'s is similar to that discussed in Section 5.3:

$$\tilde{Y}_i | x_i, s_i = j, \beta^* \overset{ind}{\sim} N(\underline{X}_i \beta_j^*, 1),$$

where $\beta_j^* \overset{i.i.d.}{\sim} N(\beta_0, C^{-1})$, for $j = 1, \ldots, k$, and the prior of the partition is given by the restricted random partition model in Section 5.3.

Simple calculations show that given the partition, the latent $(\tilde{Y}_i)$ are independent across clusters and have a multivariate normal distribution

within cluster with parameters $\hat{y}_j^*$ and W_j^{-1},

$$f(\tilde{y}_{1:n}|x_{1:n},\rho_n) = \prod_{j=1}^{k}(2\pi)^{-n_j/2}\frac{|C|^{1/2}}{|C+\underline{X}_j^{*\prime}\underline{X}_j^*|^{1/2}} * \exp\left(-\frac{1}{2}(\tilde{y}_j^*-\hat{y}_j^*)'W_j(\tilde{y}_j^*-\hat{y}_j^*)\right),$$

where $\hat{y}_j^*$ and W_j are defined as in Section 5.2.

Further conditioning on the response, we have that

$$f(\tilde{y}_{1:n}|y_{1:n},x_{1:n},\rho_n) \propto f(\tilde{y}_{1:n}|x_{1:n},\rho_n) * \prod_{i=1}^{n}(\mathbf{1}(\tilde{y}_i>0))^{y_i}(\mathbf{1}(\tilde{y}_i\le 0))^{1-y_i}.$$

Thus, given the partition and the data, the latent $\tilde{Y}_i$'s are independent across cluster and have truncated normal distribution within cluster with parameters $\hat{y}_j^*$ and W_j^{-1} and regions defined by the observed responses.

The posterior of the partition given the data and the latent $\tilde{Y}_i$'s is

$$p(\rho_n|y_{1:n},x_{1:n},\tilde{y}_{1:n}) \propto \frac{\alpha^k}{k!}\prod_{j=1}^{k}\frac{1}{n_j} * \mathbf{1}(s_{(1)}\le\ldots\le s_{(n)}) * \prod_{j=1}^{k}\frac{|C|^{1/2}}{|C+\underline{X}_j^{*\prime}\underline{X}_j^*|^{1/2}}\exp\left(-\frac{1}{2}(\tilde{y}_j^*-\hat{y}_j^*)'W_j(\tilde{y}_j^*-\hat{y}_j^*)\right).$$

Posterior samples of the partition can be obtained based on the MCMC algorithm discussed in Section 5.4 with an added step of sampling the latent $\tilde{Y}_i$'s (see Damien and Walker [2001]).

Under the 0-1 loss function, the prediction of the response for a new subject amounts to determining

$$P(\tilde{Y}_{n+1}>0|y_{1:n},x_{1:n+1}).$$

Given ρ_{n+1} and the latent $\tilde{Y}_i$'s for the observed subjects, suppose the new subject is in cluster j, then $\tilde{Y}_{n+1}$ is normally distributed with mean $\underline{X}_{n+1}\widehat{\beta}_j$ and variance $W_{n+1,j}^{-1}$, as defined in Section 5.3.2. Thus,

$$P(\tilde{Y}_{n+1}>0|y_{1:n},x_{1:n+1},\tilde{y}_{1:n},\rho_{n+1}) = \Phi\left((\underline{X}_{n+1}\widehat{\beta}_j) * W_{n+1,j}^{1/2}\right),$$

and the predictive probability of a success for the new subject is approximated by

$$P(Y_{n+1} = 1 \mid y_{1:n}, x_{1:n+1}) \approx \sum_{s=1}^{S} \sum_{j \in \mathcal{C}(\rho_n^s)} \frac{w_j^s(x_{n+1})}{\widehat{c}_3} \Phi\left((\underline{X}_{n+1} \widehat{\beta}_j^s) * W_{n+1,j}^{1/2s} \right).$$

.

5.5.3 Extensions to multivariate data

Extending the method of Section 5.3 to handle a multivariate response is quite simple. For example, if y is continuous, one only needs to replace the local normal model with a multivariate normal model. However, extending the approach for multivariate covariates is tricky since there is no natural ordering in higher-dimensions. Here we present the general approach for enforcing a given restriction and then discuss ideas on how to determine the restriction.

For $\rho_n \in \mathcal{P}_n$, let $I_R(\rho_n)$ indicate if ρ_n satisfies some given restriction R. Recall that $\{m_i\}$ is the number of clusters of size i for $i = 1, ..., n$. Let $\mathbf{k}$ and $\{\mathbf{m}_i\}$ denote the random variables with the non-bold variables indicating the realized values, and define

$$\mathcal{P}_n(k, m_{1:n}) = \{\rho_n \in \mathcal{P}_n | \mathbf{k} = k, \mathbf{m_1} = m_1, \ldots, \mathbf{m_n} = m_n\},$$

and

$$\mathcal{P}_n^*(k, m_{1:n}) = \{\rho_n \in \mathcal{P}_n(k, m_1, \ldots, m_n) | I_R(\rho_n) = 1\}.$$

Proposition 5.5.3 *The probability measure on the random partition defined by*

$$p^*(\rho_n | x_{1:n}) = \frac{\Gamma(\alpha)\Gamma(n+1)}{\Gamma(\alpha+n)} \alpha^k \prod_{i=1}^{n} \frac{1}{i^{m_i} m_i!} * \frac{1}{|\mathcal{P}_n^*(k, m_{1:n})|} * I_R(\rho_n) \tag{5.24}$$

satisfies the constraint R and has the same marginal for k, as that induced by the Dirichlet process.

The proof that the marginal for k is the same as that of the DP is obtained by summing (5.24) over all $\rho_n \in \mathcal{P}_n(k, m_{1:n})$. The indicator function assigns zero mass to all $\rho_n \in \mathcal{P}_n(k, m_{1:n}) \setminus \mathcal{P}_n^*(k, m_{1:n})$, so that the sum may be considered only over the set $\mathcal{P}_n^*(k, m_{1:n})$. The probability is uniform for partitions in this class. Thus, multiplying by the size of $\mathcal{P}_n^*(k, m_{1:n})$, gives the marginal for $(k, m_1, \ldots, m_n)$, which is equivalent to that of the DP.

For multivariate covariates, sensible constraints could be defined by requiring that the smallest rectangles in the covariate space (or spheres or ellipsoids) containing the covariates of subjects for each cluster do not intersect. The selected shape should reflect prior belief in the regression function and the form of the regions in the covariate space in which the regression function is approximately linear. In the univariate case, restriction (5.15) can also be viewed as non-intersecting 1-dim rectangles or spheres in the covariate space. The covariate random partition of (5.16) is obtained from (5.24) by noting that the size of $\mathcal{P}_n^*(k, m_{1:n})$ is the number of unique ways to order the k cluster sizes, i.e. $k!/\prod_{i=1}^{n} m_i!$. In the multivariate case, this number will likely depend on the covariates, and a more general MCMC algorithm would need to be developed. A detailed multivariate extension will be a subject of future research.

5.6 Simulated examples

To illustrate the issues related to the large number of partitions for the DPM and joint DPM models and the implications for predictive performance, we consider three simulated data examples. The results are compared with the restricted DPM model constructed here and show how the restricted DPM model is flexible in recovering a range of regression functions.

First, we study a simple example with a piecewise linear regression function and no error, so that the two clusters are clear. A set of $n = 37$

data points were generated according to the following formulae;

$$y_i|x_i = \begin{cases} -x_i/8 + 5 & \text{if } x_i \leq 6 \\ 2x_i - 12 & \text{if } x_i > 6, \end{cases}$$
$$x_i = (0, 0.25, 0.5, \ldots, 8.75, 9).$$

The hyper-parameters are specified as follows: $\alpha = 1$, $a = 2$, $b = 1/4$,

$$\beta_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } C = \begin{bmatrix} 1/144 & 0 \\ 0 & 1/4 \end{bmatrix}.$$

To illustrate the difficulties with nonlinear regression, a simple example with a quadratic regression function is considered. For $i = 1, \ldots, 50$,

$$Y_i|x_i \overset{ind}{\sim} N(x_i^2, 1); \qquad X_i \overset{iid}{\sim} U(-5, 5).$$

The hyper-parameters are specified as follows: $\alpha = 1$, $a = 2$, $b = 1$,

$$\beta_0 = \begin{bmatrix} -12 \\ 0 \end{bmatrix} \text{ and } C = \begin{bmatrix} 1/50 & 0 \\ 0 & 1/25 \end{bmatrix}.$$

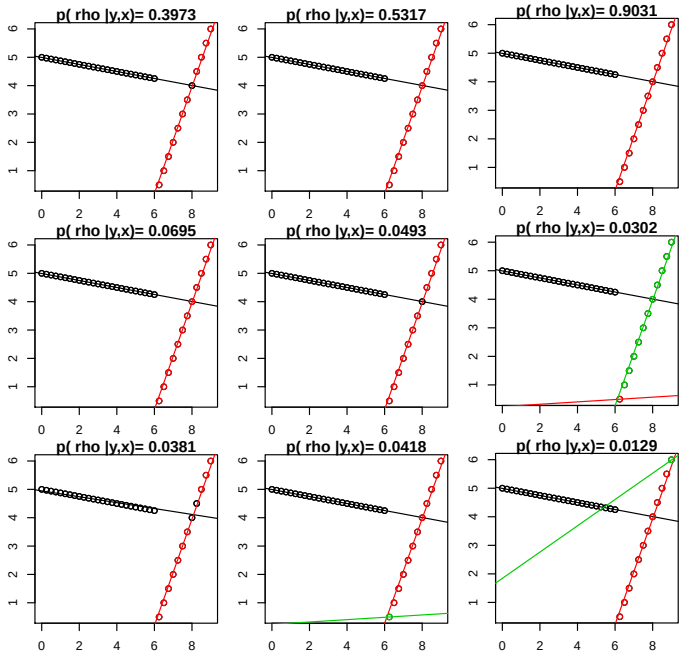
Finally, a more complicated example with $n = 100$ is generated according to

$$Y_i|x_i \overset{ind}{\sim} N(x_i \sin x_i, 1/16); \qquad X_i \overset{iid}{\sim} U(-2\pi, 2\pi).$$

The hyper–parameters are specified as follows: $\alpha = 1$, $a = 2$, $b = 1/16$,

$$\beta_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } C = \begin{bmatrix} 1/(72^2) & 0 \\ 0 & 1/144 \end{bmatrix}.$$

The MCMC scheme for the DPM model and joint DPM model (jDPM) is the Gibbs sampling method described in Neal [2000] (Algorithm 2). For the restricted DPM (rDPM) model, the algorithm described in Section 5.4 is used. All MCMC algorithms used 10,000 iterations with 1,000 burn in.



(a) DPM (b) joint DPM (c) restricted DPM

Figure 5.1: Estimated regression lines in each cluster for the three partitions with the highest estimated posterior probabilities with the data colored by cluster membership.

5.6.1 Example 1

We begin by analysing the posterior probability of the partition for the n observed subjects, since the prediction is computed based on those partitions with positive estimated probabilities.

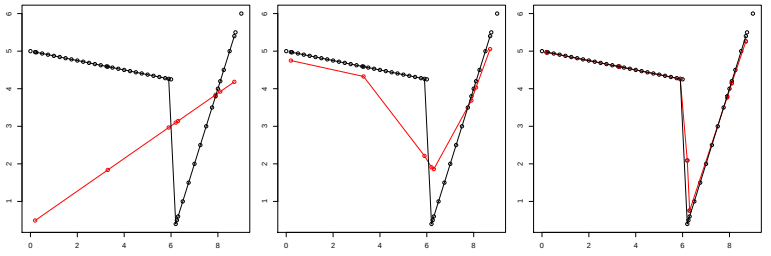
This first example demonstrates how inference for the random partition of the DPM and jDPM models can be (extremely) poor. Figure 5.1 displays the three partitions with the highest estimated probabilities for each of the models along with their corresponding probabilities. The true partition is composed of two clusters, where subjects with covariates less than 6 are in the first cluster and subjects with covariates greater than 6 are in the second cluster. The partition where the subject with a covariate of 8 is placed in the first cluster also fits the data, but is an example of undesirable partition, as too much weight will be placed on the first regression line in predictions.

The DPM model does not recognize the true partition. It gives the most weight, 0.3973, to the partition where the subject with a covariate of 8 is placed in the first cluster (in black). This occurs because more subjects are in the first cluster. Even though the correct partition has the second highest estimated probability, this value is only 0.0695.

The jDPM model is an improvement; with an estimated posterior probability of 0.5317 for the true partition, it does better at recognizing the clusters. However, the undesirable partition where the subject with a covariate of 8 is allocated to the first cluster, is still present with the second highest estimated posterior probability of 0.0493.

With an estimated posterior probability of 0.9031 for the true partition, the rDPM model is by far the best at distinguishing the clusters.

The estimated regression function at a new value of x is an average of the conditional predictions over all the 1,263 and 965 partitions with positive estimated posterior probability for the DPM and jDPM model respectively, while this average is based on only 43 partitions for rDPM model. The estimates of the regression function at $x = (0.2, 3.3, 5.9, 6.2, 6.3, 7.9, 8.1, 8.7)$ for the three models are shown in Figure 5.2. It is perhaps not surprising that the rDPM model is better at recovering the true regression function,



(a) DPM (b) joint DPM (c) restricted DPM

Figure 5.2: Prediction (in red) for $x = (3.3, 5.9, 6.2, 6.3, 7.9, 8.1, 10)$ with the true prediction (in black) and observed data (in black circles).

but it is interesting to examine what happens in the other models.

Apart from the subject with a covariate of 6.2, the cluster allocation of the new subjects is clear; the subjects with covariates of (0.2, 3.3, 5.9) should be placed in the first cluster and the subjects with covariates of (6.3, 7.9, 8.1, 8.7) should be placed in the second cluster. However, even conditionally on the true partition, the DPM and jDPM models give positive weight to the allocation of these subjects to the opposite cluster. This causes an unnecessary averaging of cluster-specific predictions across clusters that is evident in Figures 5.2a and 5.2b. For partitions other than the true one, the conditional prediction is necessarily worse. For example, consider the partition where the subject with a covariate of 8 is allocated to the first cluster. For the DPM model, the conditional prediction for new subjects in the second cluster will be overly influenced by the first regression line due to the extra individual allocated to the first cluster. For the jDPM model, the weight of first regression line will be even further inflated, especially for subjects with covariates of (7.9, 8.1), due to similarity with the covariate of 8 that is allocated to the first cluster. Allowing this partition to have positive posterior weight further contributes to the unnecessary averaging of cluster-specific predictions across clusters in Figures 5.2a and 5.2b.

By placing zero prior mass on undesirable partitions, we ensure that conditional prediction is just based on neighbouring clusters and the conditional predictions based on undesirable partitions have no impact. The prediction is greatly improved (Figure 5.2c).

We compare the empirical L_2 prediction error between the estimated prediction and the true prediction, defined by $(1/m \sum_{j=1}^{m} (\widehat{y}_{n+j,est} - \widehat{y}_{n+j,true})^2)^{1/2}$. The rDPM model, as is evident in Figure 5.2, has the smallest prediction error of 0.6029, and the jDPM and DPM models take second and third place, respectively, with prediction errors of 1.0216 and 2.3617.

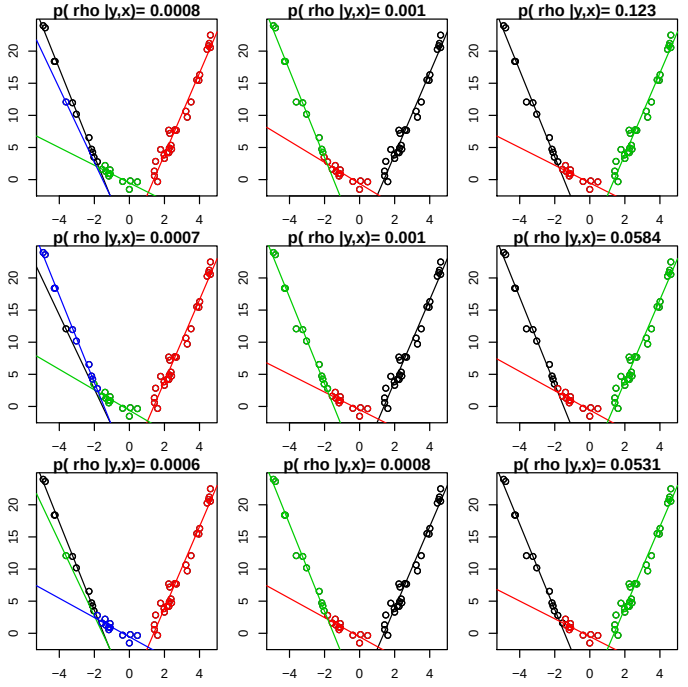
5.6.2 Example 2

In the second example, the regression curve is a quadratic function. Of course, a preliminary analysis of the plot of the data would suggest the use of a simple linear regression model with x^2 among the regressors. But, our aim here is to compare the performance of the models with this fairly well behaved curve. The three partitions with the highest estimated probabilities for the three models are depicted in Figure 5.3.

In this example, the posterior mass for the DPM and jDPM models is spread out across many partitions. In particular, for the DPM model, with 10,000 iterations, after discarding the first 1,000, a total of 9,946 partitions are visited by the chain, and for the jDPM model, this number is 9,834. Moreover the total mass of the top three partitions is only 0.0021 for the DPM model and is 0.0023 for the jDPM model.

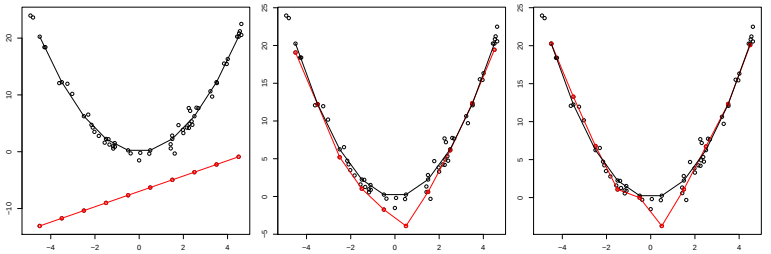
With a total of 1,044 partitions with positive estimated posterior probability and a total mass of 0.2345 for the top three partitions, the posterior mass for rDPM model is much less spread out.

The estimated regression for x from -4.5 to 4.5 by unit of 1 for the three models is displayed in Figure 5.4. The prediction for the DPM model does not even interpolate the data, and while poor prediction for this dataset was expected, the results in Figure 5.4a can appear very surprising. We emphasize that these results are due to the model. In particular, to fit the data, the clusters are associated to regions of the covariate space, and



(a) DPM (b) joint DPM (c) restricted DPM

Figure 5.3: Estimated regression lines in each cluster for the three partitions with the highest estimated posterior probabilities with the data colored by cluster membership.



(a) DPM (b) joint DPM (c) restricted DPM

Figure 5.4: Prediction (in red) for x from -4.5 to 4.5 by unit of 1 with the true prediction (in black) and observed data (in black circles).

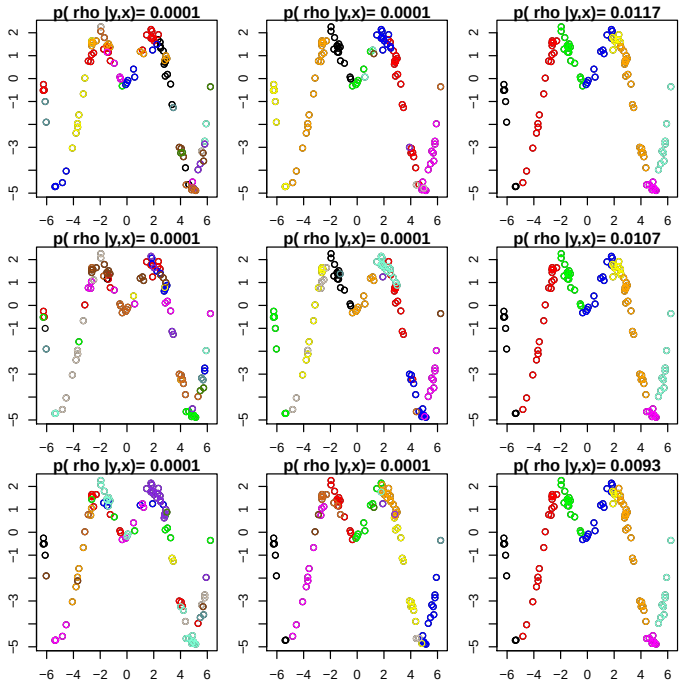
the cluster-specific predictions are averaged regardless of the value of x_{n+1} and the location of the clusters in the covariate space. This is of course an extreme example, but it does demonstrate how dramatically poor the prediction can be for the DPM model when the true regression function is nonlinear, suggesting that the DPM model should be used with caution if there is any doubt in the linearity of regression function.

Prediction for the jDPM model (Figure 5.4b) is much better but is pulled down in some regions due to the influence of predictions based on clusters in other parts of the covariate space. The prediction of the rDPM model is close to the truth for all subjects except for the subject with a covariate of 0.5 due to lack of data in that area.

Again, the rDPM model has the lowest empirical L_2 prediction error of 1.4214, while the prediction error for jDPM and DPM models are 1.6903 and 17.3154, respectively.

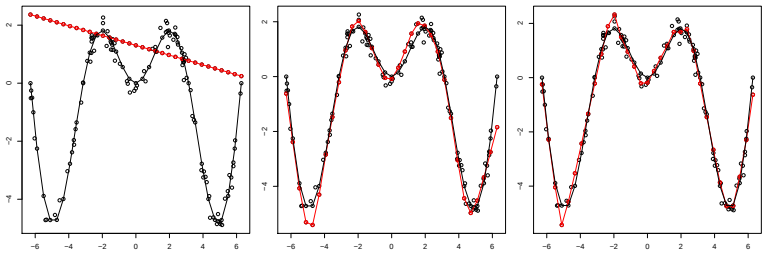
5.6.3 Example 3

The last example considers a rapidly changing regression function. This function requires many clusters for a good approximation. The three partitions with the highest estimated probabilities for the three models are depicted in Figure 5.5.



(a) DPM (b) joint DPM (c) restricted DPM

Figure 5.5: The three partitions with the highest estimated posterior probabilities colored by cluster membership.



(a) DPM (b) joint DPM (c) restricted DPM

Figure 5.6: Prediction (in red) for $m = 33$ new values of x from -2π to 2π by a unit of $\pi/8$ with the true prediction (in black) and the observed data (in black circles).

This example demonstrates how dramatically spread out the posterior for the partition can be for the DPM and jDPM models. No partitions are visited more than once for both the DPM and jDPM models. Thus, all 10,000 partitions have the same estimated posterior probability, and Figures 5.5a and 5.5b display three of them. These partitions are composed of many clusters, with an average number of clusters of 15 for the DPM model and 13 for the jDPM model. Of the partitions displayed in Figures 5.5a and 5.5b, most contain undesirable features. Nevertheless, all of these partitions are used for prediction.

For the rDPM model, on the other hand, the posterior mass is much less spread out. A total of 1,480 partitions have a positive estimated posterior probability. All partitions require at least six clusters, where the majority, 86%, of partitions have between 7 and 9 clusters.

Figure 5.6 displays the prediction for x from -2π to 2π by a unit of $\pi/8$. The DPM model again gives a linear prediction and thus, cannot capture the nonlinear regression function. For the jDPM model, the prediction is not able to react to local changes in the derivative of the curve as well as the rDPM model because it is overly influenced by data in distant regions of the covariate space. The rDPM model has the lowest empirical L_2

prediction error of 0.2578, where the prediction error for the jDPM and DPM models are, respectively, 0.4352 and 3.2762.

5.7 Alzheimer's disease study

The hippocampus is a brain structure that is relatively easily to identify and is known to affected by Alzheimer's disease. It is one of the most common neuroimaging biomarkers used to aid diagnosis of AD, but few studies have examined the extent of asymmetrical tissue loss of the left hippocampus and the right hippocampus in AD patients. This is the aim of this study, and to achieve this aim, we examine the relationship between the ratio of the volume of the left to right hippocampus and AD. Classic logistic or probit regression methods would be unable to capture the non-monotone relationship present in the data. Therefore, we use the model developed here to address this issue. In particular, we apply the rDPM model discussed in Section 5.5.2 to estimate the curve representing the probability of disease status based on the ratio of the volume of the left to right hippocampus.

The ADNI dataset analysed here consists of the volume of the left and right hippocampus obtained from the structural Magnetic Resonance Image performed at the first visit for 377 patients, of which 159 have been diagnosed with AD and 218 are cognitively normal (CN).

Let $y = 1$ indicate a healthy subject and $y = 0$ indicate a subject with AD. The covariate x represents the ratio of the volume of the left to right hippocampus. The model can be stated as follows:

$$Y_i|\beta^*, s_i = j, x_i \overset{ind}{\sim} \text{Bern}(\Phi(\underline{X}_i\beta_j^*)),$$

where

$$\beta_j^* \overset{i.i.d.}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 40 & 0 \\ 0 & 40 \end{bmatrix}\right),$$

for $j = 1, \ldots, k$, and the prior of the partition is given by the restricted random partition model in Section 5.3 with $\alpha = 1$.

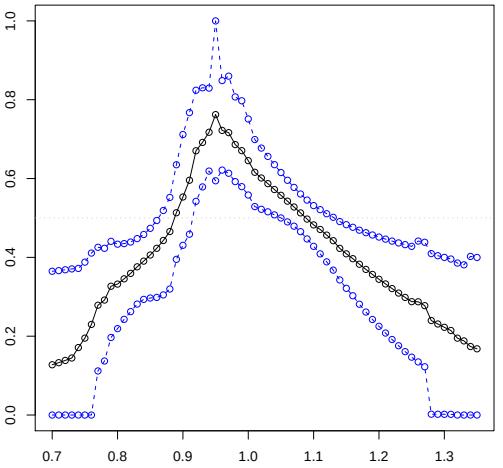


Figure 5.7: The estimated probability of being healthy (in black) for left-to-right hippocampus ratios of 0.7 to 1.35 by 0.01 with 90% credible intervals (in blue).

The algorithm discussed in Sections 5.4 and 5.5.2 with 20,000 iterations and 2,000 burn in was used to predict the probability of disease for new subjects with covariates of $x = 0.7$ to $x = 1.35$ by an interval of 0.01. Figure 5.7 displays the estimated curve with 90% pointwise credible intervals computed from the output of the MCMC. The results show the presence of asymmetrical hippocampal volume in AD patients.

Under the 0-1 loss function, patients are classified as healthy if the estimated probability is greater than 0.5; new subjects whose left hippocampus is more than 11% smaller or more than 10% larger than the right hippocampus are classified as sick. When the left hippocampus is more than 13% smaller than the right hippocampus the patient is classified as sick with at least 95% probability. This is comparable with the findings of Shi et al. [2009], who report a significant "left-less-than-right" hippocampal asymmetry pattern. However, our results also show that a "right-less-than left" hippocampal asymmetry pattern is present. In par-

ticular, the patient is classified as sick with at least 95% probability when the right hippocampus is more than 12% smaller than the left hippocampus.

5.8 Discussion

In this chapter, we provided a simple comparison of Bayesian nonparametric mixture models with constant versus covariate dependent weight functions for estimation of the regression curve, and identified a basic, but quite underestimated, problem that is present in both models.

In terms of comparison, our results demonstrate an important drawback of the model with constant weight functions and linear mean functions; it is not robust to non-linearities in regression function and can result in extremely poor prediction if non-linearity is present. This is due to the fact that inflexibility of the mean functions causes the clusters to be associated to regions of the covariate space. The local cluster-specific predictions from different parts of the covariate space are averaged together independent of x_{n+1}, resulting in poor prediction. To avoid this problem, single-p DDP models should use flexible mean functions that guarantee the regression curve described by the data can be captured by a single mean function. However, if the mean functions are too flexible, the prediction will also suffer. On the other hand, we have shown that the models with covariate dependent weight functions result in improved prediction, due to the incorporation of covariate proximity in the partition structure.

However, for both models problems arise due to the huge dimension of the partition space. In particular, the posterior puts too small a mass on desirable clusterings and too large a mass on undesirable partitions. Furthermore, an MCMC output may never even visit a partition with a desirable clustering. This occurs because it is not possible to manipulate the prior mass on partitions sufficiently, due to the extraordinarily large number of partitions and hence the microscopic probabilities involved. To address these issues, the prior knowledge on what are sensible configurations for the problem at hand needs to be introduced with extreme care.

In fact, it is appropriate to rigidly restrict the support of the prior on the random partition to the set of sensible configurations, as this is the only sure way to guarantee prominence of desirable partitions in the posterior.

To make our point, we have focused on the particular case of simple regression, with a one-dimensional covariate. When the aim is estimation of the regression function, we find it essential to assume that clusters are based on covariate proximity. We have shown the importance of highlighting these clusters in the model by putting zero weight on the alternatives. The problems of not doing this, especially poor predictive performance, have been made evident through computations and a number of examples in the chapter. We have demonstrated that the restricted DPM model is able to recover a wide range of regression functions, including functions with discontinuities, well-behaved curves, and rapidly changing curves. For other applications, the type of clustering appropriate for the data or aim must be established, and once this is understood, undesirable partitions according to the notion of clustering established should be removed. We have developed a general approach for this given the restriction, but future work is needed to examine types of constraints for regression with multivariate covariates and a suitable MCMC algorithm for inference.

Chapter 6

Normalized covariate-dependent weights

In this chapter, we discuss Bayesian nonparametric mixture models with covariate-dependent weights. The defined form of covariate-dependent weight has important implications on prediction. Thus, it is important that the weights are defined in an interpretable fashion. The various proposals in literature for direct construction of the covariate-dependent weights are based on a stick breaking representation and lack the desired property of interpretation. Moreover, extensions for inclusion of both continuous and discrete covariates are not always straightforward. Our aim in this chapter is to construct interpretable covariate-dependent weights which allow for inference with combinations of both continuous and discrete covariates. The proposed normalized weights are discussed in detail, and a novel MCMC algorithm is developed to deal with the normalizing constant. Finally, the novel model and algorithm are applied to study the evolution of one of the most widely studied AD biomarkers, hippocampal volume, as a function of age, sex, and disease status.

This chapter contains joint work with Isadora Antoniano Villalobos and Stephen G. Walker and is based on Antoniano Villalobos et al. [2012].

6.1 Introduction

The general form of Bayesian nonparametric mixture models with covariate dependent weights is

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y; x, \tilde{\theta}_j), \tag{6.1}$$

where P_x is a realization of the covariate-dependent random probability measure

$$\mathbf{P}_x = \sum_{j=1}^{\infty} w_j(x) \delta_{\tilde{\theta}_j}.$$

More generally, the atoms $(\tilde{\theta}_j)$ may also depend on x, but to simplify computations and ease interpretation, this is usually not assumed.

The main constraint when defining $(w_j(x))$ is the need to specify a prior such that

$$\sum_{j=1}^{\infty} w_j(x) = 1 \text{ a.s. for all } x \in \mathcal{X},$$

which is non trivial for an infinite number of positive weights. The popular solution, introduced by MacEachern [1999], is to define the covariate-dependent weights through the stick-breaking construction

$$w_1(x) = v_1(x),$$

$$w_j(x) = v_j(x) \prod_{j'<j} (1 - v_{j'}(x)) \quad \text{for } j > 1,$$

with the $(v_j(\cdot))$ being independent processes on $\mathcal{X}$. A wide range of models present in the literature follow this construction and differ only in the definition of the $v_j(x)$. Popular proposals include Griffin and Steele [2006], Dunson and Park [2008], Rodriguez and Dunson [2011], Chung and Dunson [2009], Ren et al. [2011] and a review is provided in Section 2.3.4.

The advantage of the stick-breaking construction is the availability of methods for exact posterior sampling. However, this construction poses other challenges.

In general, for any definition of $w_j(x)$, the weights play an important role in prediction. The prediction and predictive density are (respectively)

$$\mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}] = \int_{\mathcal{M}(\Theta)} \mathrm{E}[Y_{n+1}|x_{n+1}, P_{x_{n+1}}] dQ(P_{x_{n+1}}|y_{1:n}, x_{1:n}),$$
$$f(y|y_{1:n}, x_{1:n+1}) = \int_{\mathcal{M}(\Theta)} f(y|x_{n+1}, P_{x_{n+1}}) dQ(P_{x_{n+1}}|y_{1:n}, x_{1:n}),$$

where, assuming $K(y; x, \theta) = N(y; \underline{X}\beta, \sigma^2)$, the term inside the integral is

$$\mathrm{E}[Y_{n+1}|x_{n+1}, P_{x_{n+1}}] = \sum_{j=1}^{\infty} w_j(x) \underline{X}_{n+1} \tilde{\beta}_j,$$
$$f(y|x_{n+1}, P_{x_{n+1}}) = \sum_{j=1}^{\infty} w_j(x) N(y; \underline{X}_{n+1}\tilde{\beta}_j, \tilde{\sigma}_j^2).$$

Thus, $w_j(x)$ is the weight assigned to the local linear prediction of the j^{th} component at covariate value x and is the key for good approximation of nonlinear regression functions and complex conditional densities. Given the importance of the weights, one should have a clear understanding of the behavior of $w_j(x)$ for the chosen definition. Unfortunately, due to the nature of the stick breaking construction, a precise interpretation of how $w_j(x)$ changes with x is difficult, particularly as j increases. This makes decisions regarding the various modelling choices of $v_j(x)$, such as functional shapes and hyper-parameters, challenging, and as discussed in Section 2.3.4, the number of model choices for $v_j(x)$ is indeed quite large. Moreover, combining continuous and discrete covariates in a flexible fashion is far from straightforward.

In this chapter, we introduce an alternative construction through normalization. The normalized weights are given by

$$w_j(x) = \frac{w_j K(x; \tilde{\psi}_j)}{\sum_{j'=1}^{\infty} w_{j'} K(x; \tilde{\psi}_{j'})}, \tag{6.2}$$

where the denominator must be finite a.s. We argue in this chapter that this construction is naturally motivated in the Bayesian setting, leading to a clear understanding of behavior of the weights and allowing a simple choice of the kernel and hyperpriors involved. Moreover, it is shown to be applicable to both continuous and discrete covariates.

It is to be noted that the infinite sum in the denominator of (6.2) introduces an intractable normalizing constant for which no posterior simulation methods are available. Simulation methods are available only for the finite versions of this type of model (see e.g. Pettitt et al. [2003], Møller et al. [2006], Murray et al. [2006], Adams et al. [2009]). For this reason, only finite versions have been introduced in the literature. A further aspect of the chapter is to construct an algorithm, based on the introduction of latent variables, that solves the infinite dimensional intractable normalizing constant problem.

6.2 Regression model with normalized weights

The aim in this section is to motivate the normalization approach to the construction of weights $w_j(x)$, rather than the stick breaking construction. The idea is to associate each parametric regression model, used as a component in the mixture model, with a function that reflects where in the covariate space it applies. This results in a clear understanding of the behavior of the weights.

In the nonparametric mixture model

$$f_P(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y; x, \tilde{\theta}_j),$$

each covariate dependent weight $w_j(x)$ represents the probability that an observation with a covariate value of x comes from the j^{th} parametric regression model $K(y; x, \tilde{\theta}_j)$. Thus, letting $\tilde{s}$ be the random variable indicating the component from which the observation is generated, we have that $w_j(x) = p(\tilde{s} = j|x)$. A simple Bayes argument, implies

$$p(\tilde{s} = j|x) \propto p(\tilde{s} = j)p(x|\tilde{s} = j),$$

where $p(\tilde{s} = j)$ represents the probability that an observation comes from parametric regression model j (with the covariate of the observation unknown), and $p(x|\tilde{s} = j)$ describes how likely it is that an observation generated from regression model j has a covariate value of x.

A realistic assumption is that the parametric regression models only apply locally. In this case, $p(x|\tilde{s} = j)$ can be defined to reflect prior belief as to where in the covariate space the regression model j will provide the best description of the data. A natural way to achieve this is to define $p(x|\tilde{s} = j)$ through a parametric kernel function $K(x; \tilde{\psi}_j)$. The term $w_j = p(\tilde{s} = j)$ may penalize $K(x; \tilde{\psi}_j)$ across the covariate space, and together w_j and $K(x; \tilde{\psi}_j)$ reflect where in the covariate space regression model j applies. If the term w_j is very small, it is unlikely that regression model j will fit the data in any region of the covariate space.

Putting these things together, we have that

$$w_j(x) \propto w_j K(x; \tilde{\psi}_j),$$

and therefore, that

$$w_j(x) = \frac{w_j \, K(x; \tilde{\psi}_j)}{\sum_{j'=1}^{\infty} w_{j'} \, K(x; \tilde{\psi}_{j'})},$$

where $0 \leq w_j \leq 1$ for all j and $\sum_{j=1}^{\infty} w_j = 1$.

The key element left to define is the kernel $K(x; \tilde{\psi}_j)$. If x is a continuous covariate, a natural choice is the normal density function. In this case, the interpretation would be that there is some central location $\tilde{\mu}_j \in \mathcal{X}$ where regression model j best fits the data, and a parameter $\tilde{\tau}_j$ describing the rate at which the applicability of the model decays around $\tilde{\mu}_j$. In general, the kernel $K(x; \tilde{\psi}_j)$ may be modelled via any standard family of distribution functions. As another example, if x is discrete, then a standard distribution on discrete spaces can be used, such as the Bernoulli or its generalization, the categorical distribution. In the Bernoulli case, a parameter $\tilde{\rho}_{1,j}$ describes the probability that the given regression model j best applies at $x = 0$ and $\tilde{\rho}_{2,j} = 1 - \tilde{\rho}_{1,j}$ describes the probability that it best applies at $x = 1$. Even if x is a combination of both discrete

and continuous covariates, it is still possible to specify a joint density by combining both discrete and continuous distributions. This will be explained and demonstrated in later sections.

6.3 Latent model

Given a sample $\{(y_1, x_1), \ldots, (y_n, x_n)\}$, the likelihood function for the model with normalized weights is given by

$$f_P(y_{1:n} \,|\, x_{1:n}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{\infty} w_j(x_i)\, K(y_i; x_i, \tilde{\theta}_j) \right),$$

with covariate dependent weights given by

$$w_j(x) = \frac{w_j\, K(x; \tilde{\psi}_j)}{\sum_{j=1}^{\infty} w_j\, K(x; \tilde{\psi}_j)}.$$

The expression in the denominator can be seen as an intractable normalizing constant. In this section, we show how to undertake Bayesian inference for this model by extending the likelihood to obtain a viable latent model. We rely on a simple series expansion,

$$\sum_{k=0}^{\infty} (1-r)^k = r^{-1}, \text{ for } 0 < r < 1, \tag{6.3}$$

as the key for incorporating auxiliary variables to the likelihood expression, thus obtaining a viable latent model.

In order to illustrate ideas with a simplified notation, we start by considering posterior estimation with a single data point. The local parametric regression model is defined to be the standard linear regression model

$$K(y; x, \tilde{\theta}_j) = \mathrm{N}(y; \underline{X}\tilde{\beta}_j, \tilde{\sigma}_j^2),$$

where $\tilde{\theta}_j = (\tilde{\beta}_j, \tilde{\sigma}_j)$ and $\underline{X} = (1, x')$. We assume the first q elements of x represent discrete covariates, each x_h taking values in $\{0, \ldots, G_h\}$, for

$h = 1 \ldots, q$; the last p elements of x represent continuous covariates. In this case, we define

$$K(x; \tilde{\psi}_j) = \prod_{h=1}^{q} \mathrm{Cat}(x_h; \tilde{\rho}_{j,h}) \prod_{h=1}^{p} \mathrm{N}(x_{h+q}; \tilde{\mu}_{j,h}, \tilde{\tau}_{j,h}^{-1}),$$

where $\tilde{\psi}_j = (\tilde{\rho}_j, \tilde{\mu}_j, \tilde{\tau}_j)$ and $\mathrm{Cat}(\cdot; \tilde{\rho}_h)$ represents the categorical distribution;

$$\mathrm{Cat}(x_h; \tilde{\rho}_h) = \prod_{g=0}^{G_h} \tilde{\rho}_{h,g}^{\mathbf{1}(x_h = g)}.$$

For the rest of this chapter, we will simplify the expression by assuming $\tilde{\tau}_j = \tilde{\tau}$ for all j.

The likelihood for this model may be written as

$$f_P(y|x) = \frac{1}{r(x)} \sum_{j=1}^{\infty} w_j K(x; \tilde{\psi}_j)\, K(y; x, \tilde{\theta}_j), \tag{6.4}$$

where

$$r(x) = \sum_{j=1}^{\infty} w_j\, K(x; \tilde{\psi}_j),$$

$$K(x; \tilde{\psi}_j) = \prod_{h=1}^{q+p} K(x_h; \tilde{\psi}_{j,h}),$$

and

$$K(x_h; \tilde{\psi}_{j,h}) = \begin{cases} \prod_{g=0}^{G_h} \tilde{\rho}_{h,g}^{\mathbf{1}(x_h = g)} & h = 1, \ldots, q \\ \exp\{-\frac{1}{2}\tilde{\tau}_{h-q}(x_h - \tilde{\mu}_{j,h-q})^2\} & h = q+1, \ldots, q+p. \end{cases}$$

Notice that we have redefined the kernel function $K(x; \tilde{\psi}_j)$ by cancelling the precision term $\tilde{\tau}$ from the normal density, which appears both in the numerator and the denominator of the normalized weights expression. In this way, we guarantee that $0 < r(x) < 1$ for all $x \in \mathcal{X}$, so we can apply the series expansion (6.3) to write

$$\frac{1}{r(x)} = \sum_{k=0}^{\infty} \left[1 - \sum_{j=1}^{\infty} w_j\, K(x; \tilde{\psi}_j)\right]^k = \sum_{k=0}^{\infty} \left[\sum_{j=1}^{\infty} w_j (1 - K(x; \tilde{\psi}_j))\right]^k.$$

The assumption of $\tilde{\tau}_j = \tau$ for all j allowed the precision term to cancel, ensuring $0 < r(x) < 1$. However, this assumption may be removed with mild conditions on $\tau_{j,h}$; in particular, we must constrain $\tau_{j,h} < M_h$ for some positive constants M_h for $h = 1, \ldots, p$. Computations become slightly more complicated. So for explanation purposes, we keep the assumption of $\tilde{\tau}_j = \tau$ for all j in this text.

To deal with the infinite sum over k, we consider k as a latent variable, obtaining the latent model

$$f_P(y, k|x) = \sum_{j=1}^{\infty} w_j K(x; \tilde{\psi}_j)\, K(y; x, \tilde{\theta}_j) \left[\sum_{j=1}^{\infty} w_j (1 - K(x; \tilde{\psi}_j))\right]^k.$$

After moving the infinite sum from the denominator to the numerator, we can now deal with the mixture in the usual way. In particular, the infinite sum over j can removed by introducing a latent variable $d \in \mathbb{N}$, which indicates the mixture component to which a given observation is associated. Then, we obtain

$$f_P(y, k, d|x) = w_d K(x; \tilde{\psi}_d)\, K(y; x, \tilde{\theta}_d) \left[\sum_{j=1}^{\infty} w_j (1 - K(x; \tilde{\psi}_j))\right]^k.$$

For the remaining sum, we have the exponent k to consider. We first write this term as the product of k identical terms

$$\left[\sum_{j=1}^{\infty} w_j (1 - K(x; \tilde{\psi}_j))\right]^k = \prod_{l=1}^{k} \left[\sum_{j_l=1}^{\infty} w_{j_l} (1 - K(x; \tilde{\psi}_{j_l}))\right].$$

We can then introduce k latent variables, $D_1, \ldots, D_k$, where $D_l \in \mathbb{N}$, arriving at the full latent model

$$f_P(y, k, d, D|x) = w_d K(x; \tilde{\psi}_d)\, K(y; x, \tilde{\theta}_d) \prod_{l=1}^{k} w_{D_l} (1 - K(x; \tilde{\psi}_{D_l})).$$

It is easy to check that the original likelihood (6.4) is recovered by marginalizing over the variables d, k and $D = (D_1, \ldots, D_k)$.

For a sample of size $n \geq 1$ we simply need n copies of the latent variables. Therefore, the full latent model is given by

$$f_P(y_{1:n}, k_{1:n}, d_{1:n}, D_{1:n}|x_{1:n}) = \prod_{i=1}^{n} w_{d_i} K(x_i; \tilde{\psi}_{d_i}) K(y_i; x_i, \tilde{\theta}_{d_i}) \prod_{l=1}^{k_i} w_{D_{l,i}} \left(1 - K(x_i; \tilde{\psi}_{D_{l,i}})\right).$$

Inference can be achieved via posterior simulation using the slice sampling method of Kalli et al. [2011], to deal with the infinite choices for $d_{1:n}$ and $D_{1:n}$.

Once again, the original likelihood

$$f_P(y_{1:n} \,|\, x_{1:n}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{\infty} w(x_i; \tilde{\psi}_j)\, K(y_i; x_i, \tilde{\theta}_j) \right).$$

can be easily recovered by marginalizing over the variables $d_{1:n}, k_{1:n}$, and $D_{1:n}$. However, the introduction of these latent variables makes Bayesian inference possible, via posterior simulation of the $\{w_j\}$, $\{\tilde{\theta}_j\}$ and $\{\tilde{\psi}_j\}$, as we show in the next section.

6.4 Computations

Before describing the MCMC algorithm, we must first specify the prior of $\mathbf{P}$, which is defined by a prior specification for the weights $\{w_j\}$ and parameters $\{\tilde{\theta}_j\}$ and $\{\tilde{\psi}_j\}$.

Our focus, for the prior of the weights $\{w_j\}$ is on stick-breaking priors (Ishwaran and James [2001]). For some positive sequence of parameters $\{\zeta_{1,j}, \zeta_{2,j}\}_{j=1}^{\infty}$, the weights are defined by

$$v_j \overset{ind}{\sim} \text{Beta}(\zeta_{1,j}, \zeta_{2,j}),$$

$$w_1 = v_1,$$

$$w_j = v_j \prod_{j'<j} (1 - v_{j'}) \text{ for } j > 1.$$

Some important examples of this type of prior are 1) the Dirichlet process, when $\zeta_{1,j} = 1$ and $\zeta_{2,j} = \zeta$ for all j; 2) the two parameter Poisson-Dirichlet process, when $\zeta_{1,j} = 1-\zeta_1$ and $\zeta_{2,j} = \zeta_2+j\zeta_1$ for $0 \leq \zeta_1 < 1$ and $\zeta_2 > -\zeta_1$; and 3) the two parameter Stick-Breaking Process where $\zeta_{1,j} = \zeta_1$ and $\zeta_{2,j} = \zeta_2$ for all j.

To complete the prior specification, we assume the pairs $(\tilde{\theta}_j, \tilde{\psi}_j)$ are i.i.d. from some fixed distribution F_0 and independent from the $\{v_j\}$. We define F_0 through its associated density f_0, defined by the product of the following components,

$$f_0(\tilde{\beta}_j, \tilde{\sigma}_j^2) = \mathrm{N}(\tilde{\beta}_j; \beta_0, \tilde{\sigma}_j^2 C^{-1})\mathrm{Gamma}(1/\tilde{\sigma}_j^2; \alpha_1, \alpha_2);$$

$$f_0(\tilde{\tau}) = \prod_{h=1}^{p} \mathrm{Gamma}(\tilde{\tau}_h; a_h, b_h);$$

$$f_0(\tilde{\mu}_j \,|\, \tilde{\tau}) = \prod_{h=1}^{p} \mathrm{N}(\tilde{\mu}_{j,h}; \mu_{0,h}, (\tilde{\tau}_h c_h)^{-1});$$

$$f_0(\tilde{\rho}_j) = \prod_{h=1}^{q} \mathrm{Dir}(\tilde{\rho}_{j,h}; \gamma_h).$$

Together with the joint latent model, this provides a joint density for all the variables which need to be sampled for posterior estimation, i.e. the variables $\{w_j, \tilde{\theta}_j, \tilde{\psi}_j, k_i, d_i, D_{l,i}\}$.

However, there is still an issue due to the infinite choice of the $(d_i, D_{l,i})$, which we overcome through the slice sampling technique of Kalli et al. [2011]. Accordingly, in order to reduce the choices represented by $(d_i, D_{l,i})$ to a finite set, we introduce new latent variables, $(\nu_i, \nu_{l,i})$, which interact with the model through the following indicator functions

$$\mathbf{1}\left(\nu_i < e^{-\xi d_i}\right) \quad \text{and} \quad \mathbf{1}\left(\nu_{l,i} < e^{-\xi D_{l,i}}\right),$$

for some $\xi > 0$. Hence, the full conditional distributions for the index variables are given by

$$P(d_i = j | \ldots) \propto w_j e^{\xi j} K(x_i; \tilde{\psi}_j) K(y_i; x_i, \tilde{\theta}_j)\, \mathbf{1}(1 \leq j \leq J_i),$$

and

$$P(D_{l,i} = j|\ldots) \propto w_j e^{\xi j} \left(1 - K(x_i; \tilde{\psi}_j)\right) \mathbf{1}(1 \leq j \leq J_{l,i}),$$

where

$$J_i = \lfloor -\xi^{-1} \log \nu_i \rfloor; \quad J_{l,i} = \lfloor -\xi^{-1} \log \nu_{l,i} \rfloor.$$

Let $J = \max_{l,i}\{J_i, J_{l,i}\}$. At any given iteration, the full conditional densities for the variables involved in the MCMC algorithm do not depend on values beyond J, so we only need to sample a finite number of the $(\tilde{\psi}_j, \tilde{\theta}_j, w_j)$.

The $\{w_j\}_{j=1}^J$ can be updated at each iteration of the MCMC algorithm in the usual way, that is, by making $w_1 = v_1$ and, for $j > 1$, $w_j = v_j \prod_{j'<j}(1 - v_{j'})$. The $\{v_j\}$ must be independently sampled from the corresponding full conditionals, which can easily be identified as

$$f(v_j|\ldots) = \text{Beta}(\zeta_{1,j} + n_j + N_j, \zeta_{2,j} + n_j^+ + N_j^+),$$

where

$$n_j = \sum_i \mathbf{1}(d_i = j); \quad N_j = \sum_{l,i} \mathbf{1}(D_{l,i} = j);$$
$$n_j^+ = \sum_i \mathbf{1}(d_i > j); \quad N_j^+ = \sum_{l,i} \mathbf{1}(D_{l,i} > j).$$

The variables involved in the linear regression kernel, $(\tilde{\beta}_j, \tilde{\sigma}_j^2)$, are updated in the standard way, well known in the context of Bayesian regression. We sample independently for each j, from the full conditional density

$$f(\tilde{\beta}_j, \tilde{\sigma}_j^2|\ldots) = \text{N}(\tilde{\beta}_j; \hat{\beta}_j, \tilde{\sigma}_j^2 \hat{C}_j^{-1})\text{Gamma}(1/\tilde{\sigma}_j^2; \hat{\alpha}_{1j}, \hat{\alpha}_{2j}),$$

where

$$\begin{aligned}
\hat{\beta}_j &= \hat{C}_j^{-1}(C\beta_0 + \underline{X}_j^{*\prime} y_j^*); \\
\hat{C}_j &= C + \underline{X}_j^{*\prime}\underline{X}_j^*; \\
\hat{\alpha}_{1j} &= \alpha_1 + n_j/2; \\
\hat{\alpha}_{2j} &= \alpha_2 + \tfrac{1}{2}(y_j^* - \underline{X}_j^*\tilde{\beta}_0)' W_j (y_j^* - \underline{X}_j^*\tilde{\beta}_0); \\
W_j &= I_j - \underline{X}_j^* \hat{C}_j^{-1} \underline{X}_j^{*\prime}.
\end{aligned}$$

Here, $\underline{X}_j^*$ denotes the matrix with rows given by $\underline{X}_i = (1, x_i')$ for $d_i = j$; y_j^* is defined analogously; and I_j denotes the identity matrix of size n_j.

To update the $\{\tilde{\psi}_j\}_{j=1}^J$, it is convenient to introduce an additional set of latent variables. In order to do so, observe that, for any integer H and vector $(K_1, \ldots, K_H) \in (0,1)^H$, the following identity holds

$$1 - \prod_{h=1}^{H} K_h = \sum_{u \in \mathcal{U}} \int_{(0,1)^H} \prod_{h=1}^{H} \left[u_h \mathbf{1}\left(U_h < K_h\right) + (1-u_h)\mathbf{1}\left(U_h > K_h\right) \right] \mathrm{d}U,$$

where $U = (U_1, \ldots, U_H)$, $u = (u_1, \ldots, u_H)$ and $\mathcal{U} = \{0,1\}^H \setminus \{0\}^H$ is the set of H-dimensional vectors of zeros and ones with at least one zero entry.

We can, therefore, introduce latent variables $(u_{i,l,h}, U_{i,l,h})$, for $i = 1, \ldots, n$, $l = 1, \ldots, k_i$ and $h = 1, \ldots, q+p$, to deal with the terms $(1 - \prod_h K(x_{i,h}; \tilde{\psi}_{j,h}))$ in the likelihood. The full conditional density for $\{\tilde{\psi}_j\}_{j=1}^J$ is thus extended to the latent expression

$$\begin{aligned}
f(\tilde{\psi}_{1:J}, \{u_{i,l,h}\}, \{U_{i,l,h}\} | \ldots) \propto \prod_{j=1}^{J} f_0(\tilde{\psi}_j) \prod_{i=1}^{n} \prod_{h=1}^{q+p} K(x_{i,h}; \tilde{\psi}_{d_i,h}) \\
\prod_{l=1}^{k_i} \left[u_{i,l,h} \mathbf{1}\left(U_{i,l,h} < K_{i,l,h}\right) + (1-u_{i,l,h})\mathbf{1}\left(U_{i,l,h} > K_{i,l,h}\right) \right],
\end{aligned}$$

where $K_{i,l,h} = K(x_{i,h}; \tilde{\psi}_{D_{i,l},h})$, from which the original conditional density can be recovered by marginalizing over the $(u_{i,l,h}, U_{i,l,h})$.

The latent variables $(u_{i,l,h}, U_{i,l,h})$ can be sampled from their full conditional density by first observing that they are independent across $i = l, \ldots, n$ and $l = 1, \ldots, k_i$. For each i, l, the variables $u_{i,l} = (u_{i,l,1}, \ldots, u_{i,l,p+q})$

and $U_{i,l} = (U_{i,l,1}, \ldots, U_{i,l,p+q})$ can be sampled jointly by first sampling $u_{i,l}$ and then sampling $U_{i,l}$ conditional to $u_{i,l}$.

The variable $u_{i,l}$ is a $q+p$-dimensional vector of zeros and ones with at least one zero entry. There are $2^{p+q}-1$ such vectors, and for u in this set, the variable $u_{i,l}$ is updated by sampling from to the following distribution

$$P(u_{i,l} = u|\ldots) \propto \prod_{h=1}^{q+p} \left[u_h K(x_{i,h}; \tilde{\psi}_{D_{i,l},h}) + (1-u_h)(1-K(x_{i,h}; \tilde{\psi}_{D_{i,l},h}))\right].$$

Next, conditional on $u_{i,l}$, the latent variables $U_{i,l,h}$ for $h = 1, \ldots, p+q$ are independent and uniformly distributed in the region

$$\left[K(x_{i,h}; \tilde{\psi}_{D_{i,l},h})(1-u_{i,l,h}), K(x_{i,h}|\tilde{\psi}_{D_{i,l},h})^{u_{i,l,h}}\right].$$

Therefore, the additional variables do not pose a problem for posterior simulation. Furthermore, the introduction of these new variables transforms the latent term, introduced to deal with the intractable normalizing constant, into a product of truncation terms across h for each ψ_j, which is multiplied by the usual posterior density for the nonparametric mixture. Thus, posterior sampling for the $\tilde{\psi}_{j,h}$ is achieved by sampling from truncated densities independently across j and h.

We first consider the update of the $\{\tilde{\rho}_j\}_{j=1}^J$, which is achieved by sampling each $\tilde{\rho}_{j,h}$ independently from a truncated Dirichlet distribution,

$$f(\tilde{\rho}_{j,h} \mid \ldots) \propto \mathrm{Dir}(\tilde{\rho}_{j,h}; \hat{\gamma}_{j,h})\, \mathbf{1}\ (\tilde{\rho}_{j,h} \in R_{j,h}),$$

where, for $g = 0 \ldots, G_h$,

$$\hat{\gamma}_{j,h,g} = \gamma_{j,h,g} + \sum_{d_i = j} \mathbf{1}\ (x_{i,h} = g);$$

and

$$\begin{aligned}
R_{j,h} &= \left\{\tilde{\rho} \in (0,1)^{G_h} : r^-_{j,h,g} < \tilde{\rho}_g < r^+_{j,h,g}, g = 1, \ldots, G_h\right\}, \\
r^-_{j,h,g} &= \max\left\{U_{i,l,h} * \mathbf{1}\,(x_{i,h} = g) : D_{i,l} = j \text{ and } u_{i,l,h} = 1\right\}, \\
r^+_{j,h,g} &= \min\left\{U_{i,l,h}^{\mathbf{1}\,(x_{i,h}=g)} : D_{i,l} = j \text{ and } u_{i,l,h} = 0\right\}.
\end{aligned}$$

Next, we consider τ. This variable is updated by sampling each $\tilde{\tau}_h$ independently from a truncated gamma density,

$$f(\tilde{\tau}_h \mid \ldots) \propto \text{Gamma}(\tilde{\tau}_h; \hat{a}_h, \hat{b}_h)\mathbf{1}\left(T_h^- < \tilde{\tau}_h < T_h^+\right),$$

where

$$\begin{aligned}
\hat{a}_h &= a_h + J/2, \\
\hat{b}_h &= b_h + \frac{1}{2}\sum_{i=1}^{n}(x_{i,h+q} - \tilde{\mu}_{d_i,h})^2 + \frac{1}{2}c_h\sum_{j=1}^{J}(\tilde{\mu}_{j,h} - \tilde{\mu}_{0,h})^2, \\
T_h^- &= \max\left\{\frac{-2\log U_{i,l,h+q}}{(x_{i,h+q} - \tilde{\mu}_{D_{i,l},h})^2} : u_{i,l,h+q} = 0\right\}, \\
T_h^+ &= \min\left\{\frac{-2\log U_{i,l,h+q}}{(x_{i,h+q} - \tilde{\mu}_{D_{i,l},h})^2} : u_{i,l,h+q} = 1\right\}.
\end{aligned}$$

Next, we sample each $\tilde{\mu}_{j,h}$ independently from a truncated normal

$$f(\tilde{\mu}_{j,h} \mid \ldots) \propto \text{N}(\tilde{\mu}_{j,h} \mid \hat{\mu}_{j,h}, (\tilde{\tau}_h\hat{c}_{j,h})^{-1})\,\mathbf{1}\left(\tilde{\mu}_{j,h} \in \bigcap_{D_{i,l}=j} A_{i,l,h}\right),$$

where

$$\begin{aligned}
\hat{c}_{j,h} &= c_h + n_j; \\
\hat{\mu}_{j,h} &= \frac{1}{\hat{c}_{j,h}}\left(c_h\tilde{\mu}_{0,h} + \sum_{d_i=j} x_{i,h+q}\right).
\end{aligned}$$

The truncation is defined through the intervals

$$I_{i,l,h} = \left(x_{i,h+q} - \sqrt{\frac{-2\log U_{i,l,h+q}}{\tilde{\tau}_h}},\; x_{i,h+q} + \sqrt{\frac{-2\log U_{i,l,h+q}}{\tilde{\tau}_h}}\right),$$

where $A_{i,l,h} = I_{i,l,h}$ when $u_{i,l,h+p} = 1$, and $A_{i,l,h} = I_{i,j,h}^c$ when $u_{i,l,h+p} = 0$.

Finally, for the update of each k_i, we use ideas involving a version of the reversible jump MCMC (see Green [1995]), introduced by Godsill [2001]), to deal with the change of dimension in the sampling space. We start by

proposing a move from k_i to $k_i + 1$ with probability $1/2$, and accepting it with probability

$$\min\left\{1, \left(1 - K(x_i; \tilde{\psi}_{D_{i,k_i+1}})\right)\right\}.$$

The evaluation of this expression requires the sampling of the additional index D_{i,k_i+1}, and in order to ensure reversibility of the Markov chain constructed by the algorithm, we will choose $D_{i,k_i+1} = j$ with probability w_j.

Similarly, if $k_i > 0$, a move from k_i to $k_i - 1$ is proposed with probability $1/2$, and accepted with probability

$$\min\left\{1, \left(1 - K(x_i; \tilde{\psi}_{D_{i,k_i}})\right)^{-1}\right\}.$$

We have shown it is possible to perform posterior inference for the nonparametric regression model proposed, via an MCMC scheme applied to the latent model. We have successfully implemented the method in Matlab (R2012a), and present some results in the Section 6.6.

Before presenting the results, we would like to mention that after posterior samples of $\{w_j, \theta_j, \psi_j\}$ have been obtained via the algorithm detailed in this section, the prediction and predictive density can be easily estimated by

$$\mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}] \approx \sum_{s=1}^{S}\sum_{j=1}^{J^s} w_j^s(x_{n+1})\underline{X}_{n+1}\tilde{\beta}_j^s,$$

$$f(y|y_{1:n}, x_{1:n+1}) \approx \sum_{s=1}^{S}\sum_{j=1}^{J^s} w_j^s(x_{n+1})\mathrm{N}(y; \underline{X}_{n+1}\tilde{\beta}_j^s, \tilde{\sigma}_j^{2s}),$$

where

$$w_j^s(x_{n+1}) = \frac{w_j^s K(x_{n+1}; \tilde{\psi}_j^s)}{\sum_{j'=1}^{J^s} w_{j'}^s K(x_{n+1}; \tilde{\psi}_{j'}^s)},$$

and $(w_j^s, \tilde{\theta}_j^s, \tilde{\psi}_j^s)$ for $s = 1, \ldots, S$ denote the S posterior samples.

6.5 Comparison with the joint approach

It should be noted that the DP mixture model based on the joint approach, reviewed in Section 2.2 and further discussed in Chapters 4 and 5, implies the same structure for the covariate dependent weights. The important difference is that here posterior inference is based on the conditional likelihood,

$$f(\{w_j, \tilde{\theta}_j, \tilde{\psi}_j\}|y_{1:n}, x_{1:n}) \propto f_0(\{w_j, \tilde{\theta}_j, \tilde{\psi}_j\}) \prod_{i=1}^{n} f_P(y_i|x_i).$$

Whereas, for the DP mixture model of the joint approach, posterior inference is based on the joint likelihood,

$$f(\{w_j, \tilde{\theta}_j, \tilde{\psi}_j\}|y_{1:n}, x_{1:n}) \propto f_0(\{w_j, \tilde{\theta}_j, \tilde{\psi}_j\}) \prod_{i=1}^{n} f_P(y_i, x_i).$$

We are only interested in estimation of the conditional density and thus, the parameters $(w_j, \tilde{\theta}_j, \tilde{\psi}_j)$ that fit the conditional. The model developed here has the advantage that inference is carried out directly for the conditional density, reflecting our interest. In a review paper, Müller and Quintana [2004] state that the joint approach "wrongly introduces an additional factor for the marginal of x in the likelihood and thus provides only approximate inference". In fact, as discussed in Chapter 4, by including this additional factor, components will be required to fit the marginal of x, which can degrade the performance of the conditional density estimate. Consider, for example, that $f_0(y|x) = N(y; \underline{X}\beta, \sigma^2)$ and X is uniform is some region. If our aim is estimation of the conditional density of $Y|x$ with the DP mixture model based on the joint approach, several normal components will be required to approximate the uniform distribution of X even though the conditional density of $Y|x$ can be approximated with a single component. We emphasize that this occurs because we are modelling the joint distribution, when interest is only in the conditional. Since posterior inference is based only on the conditional likelihood, the model developed here is able to overcome this problem, but it still maintains the

same natural and interpretable structure for the weights of the joint DP mixture model.

6.6 Simulated examples

6.6.1 Example 1

To demonstrate the ability of the model to recover complex regressions functions with the presence of both continuous and discrete covariates, we simulate $n = 200$ data points through the following formulas,

$$\begin{aligned} X_{1,i} &\overset{iid}{\sim} \text{Bern}(0.5), \\ X_{2,i} &\overset{iid}{\sim} \text{Unif}(-5, 5), \\ Y_i | x_i &\overset{ind}{\sim} \text{N}((\mathbf{1}(x_{1,i} = 1) - \mathbf{1}(x_{1,i} = 0)) * x_{2,i}^2, 1). \end{aligned}$$

The data are depicted in Figure 6.1. This is just a toy example, and the plot of the data clearly suggests a quadratic relationship between Y and X_2 given the value of X_1. However in higher dimensions this relationship and the required number of interactions terms would not be so obvious. Here, we consider only one continuous covariate, in order to visually depict the behavior of the covariate-dependent weights.

Our model is

$$f_P(y|x) = \sum_{j=1}^{\infty} w_j(x) N(y; \underline{X}\tilde{\beta}_j, \tilde{\sigma}_j^2),$$

where

$$w_j(x) = \frac{w_j \tilde{\rho}_{j,0}^{1_{x_1=0}} \tilde{\rho}_{j,1}^{1_{x_1=1}} \exp(-\tilde{\tau}/2(x_2 - \tilde{\mu}_j)^2)}{\sum_{j'=1}^{\infty} w_{j'} \tilde{\rho}_{j',0}^{1_{x_1=0}} \tilde{\rho}_{j',1}^{1_{x_1=1}} \exp(-\tilde{\tau}/2(x_2 - \tilde{\mu}_{j'})^2)}.$$

The prior for w_j and $(\tilde{\theta}_j, \tilde{\psi}_j)$ is described in Section 6.4. The prior parameters for w_j are $\zeta_{1,j} = 1$ and $\zeta_{2,j} = 1$, corresponding to a Dirichlet process

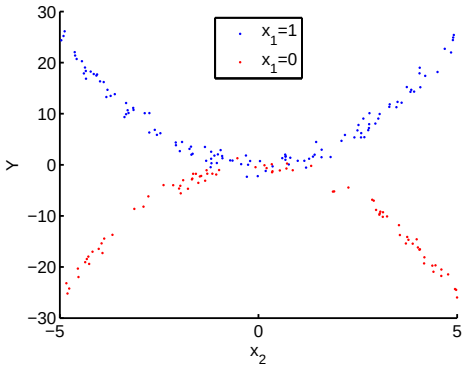


Figure 6.1: Simulated data with y plotted against x_2. The data are colored by x_1.

prior with a precision parameter of 1. For the prior of $(\tilde{\theta}_j, \tilde{\psi}_j)$, we set

$$\begin{aligned} \beta_0 = (12.5, -25, 0)'; \quad C^{-1} &= \text{diag}(50, 150, 25); \\ \alpha_1 = 1; \quad \alpha_2 &= 1; \\ \gamma = &(1, 1)'; \\ \mu_0 = 0; \quad c &= 1/4; \\ a = 1; \quad b &= 1. \end{aligned}$$

For this example, as well as for all examples presented, we explored other choices of the prior parameters including small values of a, b, c, so that the prior for $\tilde{\psi}_j$ is non-informative, and larger values for the precision parameter of the DP prior. The results were robust to these choices. Inference is carried out via the algorithm discussed in Section 6.4 with 5,000 iterations after a burn in period of 5,000. For all MCMC simulations, we examined the trace plots of the subject specific parameters. Mixing was good for all parameters, but a bit less so for $\tilde{\tau}$. We believe that an extension of the model and algorithm with component specific $\tilde{\tau}_j$ would improve the mixing, and an implementation of this algorithm is an object of current research. However, we do find that the estimates of the regression function

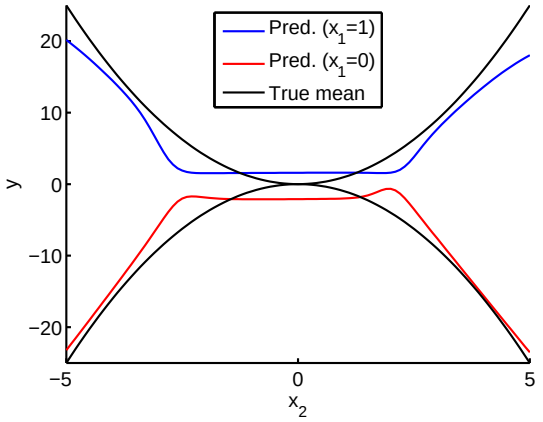
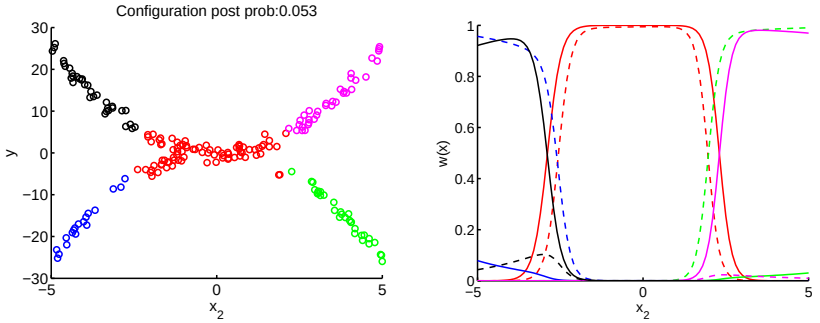


Figure 6.2: Predicted regression function for a grid of new covariate values. The black line represent the true function, while the blue and red represent the predicted function for $x_1 = 1$ and $x_1 = 0$ respectively.

and conditional density are stable to increases (or decreases) in the number of iterations and burn in period with the current algorithm.

Figure 6.2 depicts the predicted regression function for a grid of x_2 values with $x_1 = 1$ in blue and $x_1 = 0$ in red. The true regression function is shown in black. Even though the true function is quite peculiar, the model is able to recover it well.

This flexibility in estimating the regression function relies heavily on the posterior of the covariate dependent weights. The posterior of the partition is spread out among many similar partitions, and in the left panel of Figure 6.3 a representative partition, the partition with highest estimated posterior probability, is depicted with data points colored by component membership. The right panel of Figure 6.3 plots a posterior sample of the covariate-dependent weights as a function of x_2, given this partition. Solid lines denote the case when $x_1 = 1$ and dashed lines denote when $x_1 = 0$. It is important to observe that *aposteriori* the weights are able to peak close to one in areas of high applicability of their associated



(a) Partition with highest prob. (b) Covariate-dependent weights

Figure 6.3: The left panel depicts the partition with the highest posterior probability, where the data are colored by component membership. The right panel depicts the covariate-dependent weights associated to this partition with solid lines representing $w_j(1, x_2)$ and dashed lines representing $w_j(0, x_2)$ for a grid of x_2 values.

linear regression models and decay smoothly or sharply, as needed, when the covariates move away from this area.

6.6.2 Example 2

In many situations, the error distribution may also evolve with x. We consider such a situation in the following example, where $n = 200$ data points are simulated assuming a linear mean function and increasing variance;

$$X_i \overset{iid}{\sim} \text{Unif}(0, 10),$$
$$Y_i | x_i \overset{ind}{\sim} \text{N}\left(.5x_i, .25 + \exp\left(\frac{x_i - 10}{2}\right)\right).$$

Figure 6.4 displays the data.

Our model is

$$f_P(y|x) = \sum_{j=1}^{\infty} w_j(x) N(y; \underline{X}\tilde{\beta}_j, \tilde{\sigma}_j^2),$$

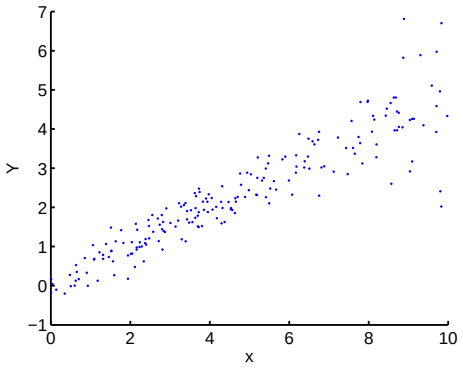


Figure 6.4: Simulated data with y plotted against x.

where

$$w_j(x) = \frac{w_j \exp(-\tilde{\tau}/2(x-\tilde{\mu}_j)^2)}{\sum_{j'=1}^{\infty} w_{j'} \exp(-\tilde{\tau}/2(x-\tilde{\mu}_{j'})^2)}.$$

The prior parameters for w_j are $\zeta_{1,j} = 1$ and $\zeta_{2,j} = 1$, and for $(\tilde{\theta}_j, \tilde{\psi}_j)$, we select

$$\begin{aligned} \beta_0 = (0, .5)'; \quad & C^{-1} = \text{diag}(10, 1/4); \\ \alpha_1 = 1; \quad & \alpha_2 = 1; \\ \mu_0 = 5; \quad & c = 1/4; \\ a = 1; \quad & b = 1. \end{aligned}$$

Inference is carried out with 5,000 iterations after a burn in period of 5,000.

Figure 6.5 depicts the predicted regression function for a grid of x values (blue solid line) and 95% pointwise credible intervals (blue dashed lines). The true regression function is shown in black. The true regression function is a simple linear function, and the model recovers it well.

Since the regression function is linear, observing Figure 6.5 could lead one to believe that all subjects belong to the same component with a high posterior probability. However, there is a more complex aspect to this

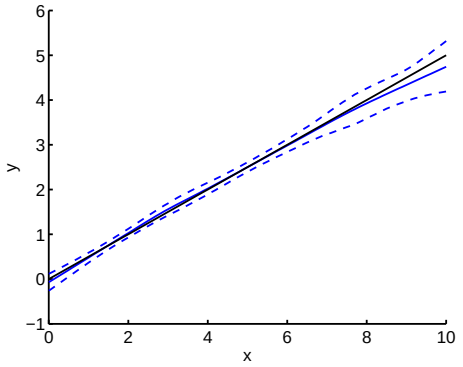


Figure 6.5: Predicted regression function for a grid of new covariate values. The black line represent the true function, while the blue represents the predicted function and the blue dashed lines provide 95% credible intervals.

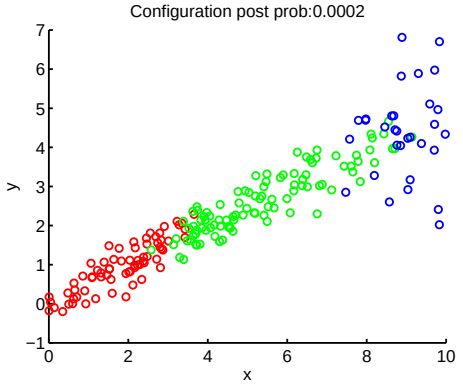


Figure 6.6: The configuration with the highest posterior probability, where the data are colored by component membership.

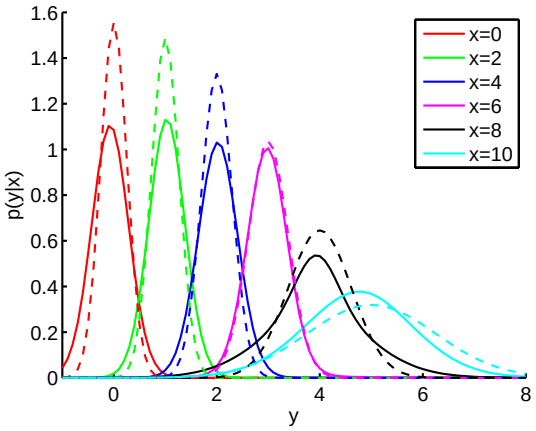


Figure 6.7: The predictive density for $x = 0, 2, 4, 6, 8, 10$ with solid lines denoting the prediction and dashed lines denoting the true density.

example; the variance of the error distribution increases with x. In fact, posterior samples of the configurations mostly consist of 3 clusters to capture this. Again, the posterior of the partition is spread out across many similar partitions, and the partition with the highest posterior probability is depicted in Figure 6.6, as a representative partition.

The predictive density at a grid of y values was estimated for all new x values. Figure 6.7 displays the predictive density for covariate values of $x = 0, 2, 4, 6, 8, 10$. The predictive density estimates across the grid of new x values are summarized by their 95% credible intervals in Figure 6.8; this provides the 95% credible intervals for the response of a new subject Y_{n+1} given x_{n+1} for a grid of new x values. Although the density at the mode and the variance are slightly underestimated and overestimated, respectfully, for small values of the covariates, the general dynamics of the variance function are well captured. Furthermore, the 95% credible intervals for $Y|x$ contain the observations and seem to accurately reflect the information present in the data.

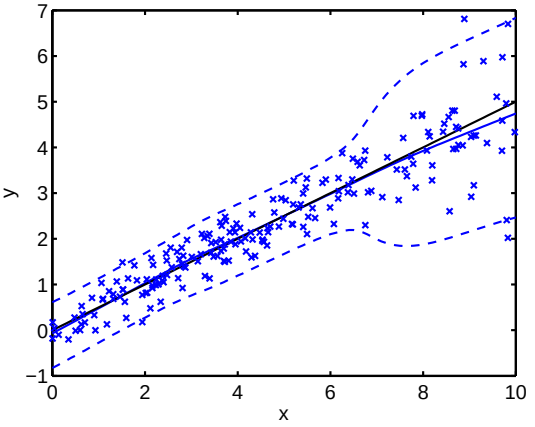


Figure 6.8: The 95% credible intervals computed from the predictive density along with the data and prediction.

6.7 Alzheimer's disease study

Understanding the dynamics of Alzheimer's disease biomarkers is important for the development of disease-modifying drugs or therapies in a clinical trial setting. In particular, those which change earliest and fastest should be used for diagnosis or as inclusion criteria for the trials; those which change the most in the disease stage of interest should be used as outcome measures for the trials; and all should be combined to assess the disease stage of the individual. In two recent papers, Jack et al. [2010] and Frisoni et al. [2010] discussed a hypothetical model for the dynamics of five well studied AD biomarkers as a function of age and disease status, including hippocampal volume.

Hippocampal volume is one of the best established and most studied biomarkers because of its known association with memory skills and relatively easy identification in sMRI. It will be our biomarker of focus for this study.

The clinical stages of the AD are divided into three phases (Jack et al. [2010]); the pre-symptomatic phase, prodromal phase, and the dementia

phase. During the pre-symptomatic phase, some AD pathological changes are present, but patients do not exhibit clinical symptoms. This phase may begin possibly 20 years before the onset of clinical symptoms. The pre-prodromal stage of AD is known as mild cognitive impairment (MCI); patients diagnosed with MCI exhibit early symptoms of cognitive impairment, but do not meet the dementia criteria. The final stage of AD is dementia, when patients are officially diagnosed AD.

Jack et al. [2010] and Frisoni et al. [2010] hypothesized that hippocampal volume evolves sigmoidally over time, with changes starting slightly before the MCI stage and occurring until late in dementia phase. The steepest changes are supposed to occur shortly after the dementia threshold has been crossed. Moreover, departures from the classical i.i.d. normality assumption of the errors are expected, due to variability in the onset of the disease and other factors, such as enhanced cognitive reserve or undiscovered neuroprotective genes.

To provide validation for this model, a flexible nonparametric model is considered to study the evolution of hippocampal volume as a function of age, gender, and disease status. The ADNI dataset analysed here consists of the volume hippocampus obtained from the sMRI performed at the first visit for 736 patients. Of the 736 patients in our study, 159 have been diagnosed with AD, 357 have MCI, and 218 are cognitively normal (CN). Figure 6.9 displays the data.

We consider the model developed in this chapter, specifically,

$$f_P(y|x) = \sum_{j=1}^{\infty} w_j(x) N(y; \underline{X}\tilde{\beta}_j, \tilde{\sigma}_j^2),$$

where

$$w_j(x) = \frac{w_j \prod_{h=1}^{2} \prod_{g=0}^{G_h} \tilde{\rho}_{j,h,g}^{1_{x_h=g}} \exp(-\tilde{\tau}/2(x_3 - \tilde{\mu}_j)^2)}{\sum_{j'=1}^{\infty} w_{j'} \prod_{h=1}^{2} \prod_{g=0}^{G_h} \tilde{\rho}_{j',h,g}^{1_{x_h=g}} \exp(-\tilde{\tau}/2(x_3 - \tilde{\mu}_{j'})^2)},$$

$G_1 = 1$ (gender) and $G_2 = 2$ (disease status). Note that here age (x_3) is a real number measuring time from birth to exam date and thus, is treated as a continuous covariate. The prior for w_j and $(\tilde{\theta}_j, \tilde{\psi}_j)$ is described in

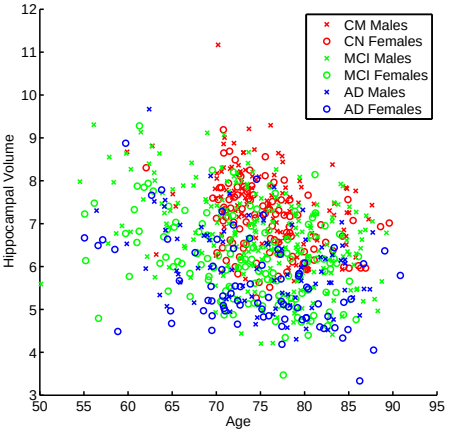


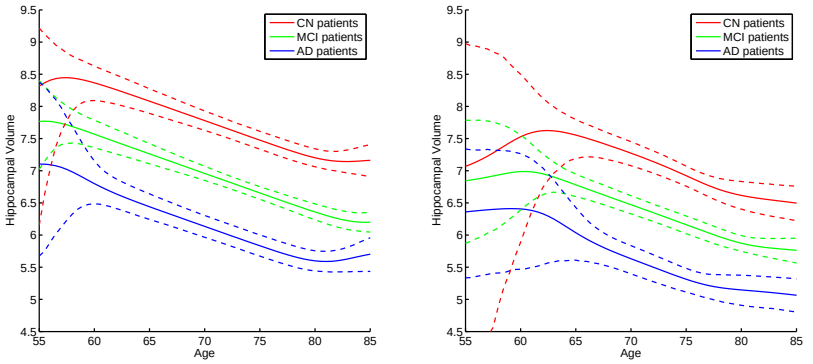
Figure 6.9: Hippocampal volume plotted against age. The data are colored by disease status with circles representing females and crosses representing males.

Section 6.4. The prior parameters for w_j are $\zeta_{1,j} = 1$ and $\zeta_{2,j} = 1$, corresponding to a Dirichlet process prior with a precision parameter of 1. For the prior of $(\tilde{\theta}_j, \tilde{\psi}_j)$, we set

$$\begin{aligned} \beta_0 = (8, -1, -1, -1/4)'; \quad & C^{-1} = \text{diag}(4, 1/4, 1/4, 1/60); \\ \alpha_1 = 1; \quad & \alpha_2 = 1; \\ \gamma_1 = (1, 1)'; \quad & \gamma_2 = (1, 1, 1)'; \\ \mu_0 = 72.5; \quad & c = 1/4; \\ a_1 = 1; \quad & b_h = 1. \end{aligned}$$

Inference is carried out via the algorithm discussed in Section 6.4 with 5,000 iterations after a burn in period of 5,000.

Figure 6.10 displays the predicted regression function for a grid of ages with all possible combinations of disease status and sex. Color indicates disease status, while results for males are displayed in the left panel and those for females are in the right panel. Interestingly, we observe a confir-



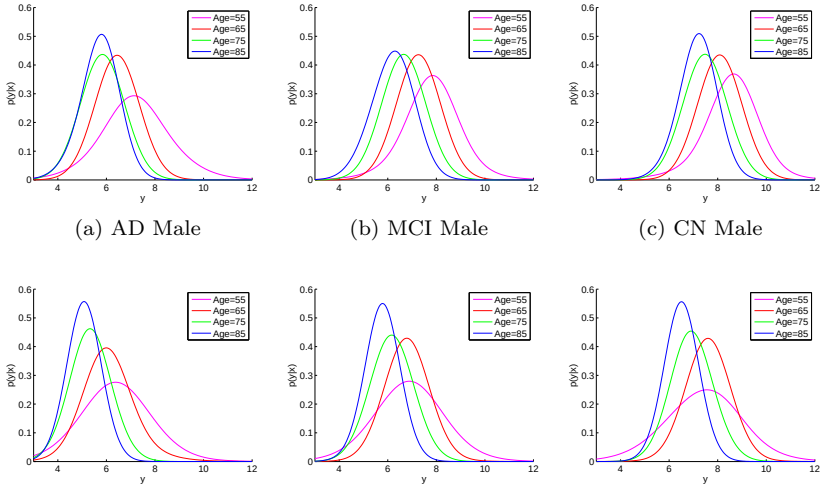
(a) Male patients

(b) Female patients

Figure 6.10: Predicted hippocampal volume as a function of age, disease, and sex. The data are colored by disease status with dashed lines representing 95% pointwise credible intervals around the predictive function.

mation of hypothesized sigmoidal evolution of hippocampal volume as a function of age. Cognitively normal subjects are predicted to have highest values of hippocampal volume at all ages, and MCI patients are predicted to have higher values of hippocampal volume at all ages when compared with AD patients. This indicates that hippocampal volume may be useful in disease staging during both the MCI and AD phases. With careful examination of Figure 6.10, we observe that the estimated curve for CN patients, as a function of age, displays the most gradual decline, while the estimated curve for AD patients displays the greatest decline. Notice that, as expected, females are predicted to have lower values of hippocampal volume, but the start of the decline in the curve has a lag of approximately five years when compared to males. We should comment that there is no data for the subgroup of CN females under 60, which reflects on the greater uncertainty in the estimation.

Figure 6.11 displays the predictive density estimates given new covariates with ages of 55, 65, 75, and 85 and all combinations of disease status



(a) AD Male (b) MCI Male (c) CN Male

(d) AD Female (e) MCI Female (f) CN Female

Figure 6.11: Conditional density estimates for new covariates with ages of 55, 65, 75, and 85 and all combinations of disease status and sex.

and sex. In a clinical trial setting, the preference is for reliable outcome measures, i.e. biomarkers with small variability. In general, we observe that variance decreases with subjects of a higher age, indicating that hippocampal volume is more reliable for elderly patients. The difference is more extreme for females as opposed to males. In particular, hippocampal volume is predicted to have a large variability for young females across all disease stages, with the largest for young CN females (the subgroup with no data). Instead, for older females, the variance is much smaller for all disease stages. When comparing males across disease status, we notice that young AD patients are predicted to show a large variability compared with young MCI and CN patients, while old MCI patients are predicted to show the largest variability when compared with their CN and AD counterparts.

6.8 Discussion

In this chapter, we have developed a novel Bayesian nonparametric regression model based on normalized covariate-dependent weights. The contribution of this construction over stick-breaking methods is the natural and interpretable structure for the weights. Other important contributions include a novel algorithm for exact posterior inference and the inclusion of both continuous and discrete covariates.

We have focused on a univariate and continuous response, but the model and algorithm can be easily extended to accommodate other types of responses by, for example, simply replacing the normal linear regression component for y with a generalized linear model. As discussed in Section 6.3, the model can also be generalized to allow multiple τ. We intend to extend the code to allow for these generalizations and make the code publicly available.

A potential downside of this approach is that computations can become intensive with large n and p. Further work is needed to examine the behavior of the model and algorithm for increasing n and p and discover potential sources of improvement in the algorithm to speed up computa-

tions while maintaining good mixing.

Additional future work will consist of examining theoretical properties of this model.

In Section 6.7, we used a fully nonparametric approach to examine the evolution of hippocampal volume as a function of age, gender, and disease status. We find that on average hippocampal volume, as a function of age, is predicted to display a sigmoidal decline for cognitively normal, MCI, and AD patients. We also observe that the decline in the curve is the most gradual for CN patients, while for AD patients, the decline is the steepest. As the approach was nonparametric, no structure was assumed for the regression function, yet our results confirm the hypothetical dynamics of hippocampal volume proposed by Jack et al. [2010]. This provides strong statistical support for their model of hippocampal volume decline.

Future work in this application, will involve examining the dynamics of various biomarkers jointly, which could be accomplished by replacing the normal linear regression component for y with a multivariate linear regression component. Another important future study will consist of combining the cross-sectional data with the longitudinal data for each patient.

Chapter 7

Discussion

Bayesian nonparametric regression mixture models are numerous and highly flexible, so that, ideally, they should be able to adapt to the behavior of Y given x present in any dataset. This raises the question of how to choose among the various models for the dataset at hand. To answer this question, practical and computational aspects of the models need to be highlighted, and a detailed study of properties needs to be carried out. This thesis was aimed at exploring some of these issues, particularly, through a detailed analysis of the prediction, but a more pragmatic comparison through computations and simulations was also explored.

Mixture models for covariate-dependent density estimation can, for the most part, be categorized into three main types of models 1) joint mixture models for (Y, X), 2) covariate-dependent mixture models with flexible mean functions and constant weights, and 3) covariate-dependent mixture models with flexible weights and linear mean functions. Both within and across model type, we have highlighted advantages and disadvantages.

For joint mixture models, the DP is selected as the mixing measure in almost all proposals because of its well known sampling procedures and desirable properties such as easy elicitation of the parameters, large support, and posterior consistency. Joint DP mixture models are computationally the easiest among the three model types, and as shown through the ex-

amples in Sections 4.5 and 5.6, perform well in practice from a predictive perspective. Thanks to the parametrization of Shahbaba and Neal [2009], extensions for various types of responses and multivariate and mixed types of covariates are straightforward.

However, a downside of this approach is that posterior inference is based on the joint likelihood, which may have undesirable effects on the conditional mean and density estimates, particularly as p increases. In Chapter 4, we focus on a typical situation in problems with high-dimensional covariates: when the marginal density of x requires many kernels for a good approximation. We carefully study the effects of using the joint likelihood in this situation and show that replacing the DP with the EDP can lead to more efficient estimators in terms of smaller estimation errors and tighter credible intervals. Moreover, computations remain quite easy for the EDP joint mixture model.

The second type of models, those with covariate-dependent mean functions and constant weights, can also be relatively simple from a computational perspective. The main modelling choice for this model type is the form of the mean functions, which, to achieve modelling flexibility, needs to be flexible. However, highly flexible mean functions can greatly increase the computational cost of the model. In fact, in Chapter 5, on the basis of careful examination of the prediction and simulated examples, we concluded that caution should be exercised when using this type of models. One may be tempted to use a simplified mean structure to ease computations, but the specified mean structure has strong implications for the estimated regression function. In particular, if the regression function present in the data cannot be well approximated by a single mean function, then (extremely) poor estimation of the regression function may result. On the other hand, an overly flexible mean function can also decrease the predictive power of the model. Thus, one must have a strong belief in the form of mean function for these types of models. Moreover, defining the appropriate mean function when multivariate and mixed types of covariates are present can be challenging.

The third, and final, model type with covariate-dependent weights

tends to be the most difficult from a computational perspective, but like the joint model, these models imply a covariate-dependent partitioning of the data, which as discussed in Chapter 5, can greatly improve prediction. However, unlike the joint model, posterior inference is based directly on the conditional likelihood of $Y|x$.

Most proposals for covariate-dependent weights in literature are constructed through a stick-breaking representation. An overlooked issue of this construction is the lack of interpretation of the covariate-dependent weights, which amplifies the difficulty in selecting the various parameters and functional shapes discussed in Section 2.3.4 that are needed to define the weights. Since flexible prediction relies heavily on the covariate-dependent weights, degraded predictive performance may also result from this. These issues can be overcome through the proposed normalized weights of Chapter 6.

In Chapter 5, we focus on estimating the regression function and carefully examine the effect of the huge dimension of the partition space, an issue common to all model types. We find that strictly enforcing the notion of covariate proximity in the partition structure can improve estimates of the regression function, but further work is needed for an extension to multivariate covariates.

In summary, we find that the joint DP mixture model is computationally the simplest but suffers from the drawback that posterior inference is based on the joint likelihood, when interest is in the conditional. The second type of model with covariate-dependent atoms overcomes this, but requires a careful balance of under and over flexibility of the mean function. Furthermore, computational complexity increases as flexibility of the mean function increases. The third type of model with covariate dependent weights also overcomes this problem, again, at some computational cost. Moreover, when the weights are constructed through normalization, this problem is overcome while maintaining the same structure for the conditional density as the joint DP mixture model and allowing simple choices of the parameters. Finally, we find the estimates can improve when prior information regarding the partition structure, such as covariate proximity,

is enforced.

High-dimensional datasets are becoming increasingly abundant. The EDP model is a simple adaptation of the joint model to deal with its shortcomings in high-dimensions. Computations for the EDP mixture model remain relatively simple. However, since the number of x-kernels is likely to be large in high-dimensions, one may be worried that computations may become burdensome for increasing p. This effect clearly depends on the dataset and further work is needed to explore it. A possible extension for future research is to combine the EDP mixture model with dimension reduction techniques.

The model based on normalized weights is methodologically attractive, but may not be well suited to large p problems for computational reasons. In particular, exact posterior sampling is available via the introduction of latent variables, but the number of latent variables is likely to increasing greatly with p. Further work is needed to explore the behavior of the model and algorithm in high-dimensions and, if needed, to develop possible extensions in this setting.

In this thesis, properties of Bayesian nonparametric regression mixture models were examined by deriving predictive equations of the conditional mean and density estimate and analysing in detail the quantities involved. This work formed the basis for a comparison of the Bayesian nonparametric models and priors of interest. A general open problem is to what extent these comparisons can be formalized. In fact, formal model comparison, in general, is a debated and underdeveloped subject in the Bayesian nonparametric community.

Formal model properties are mostly studied in terms of frequentist properties, and the first step in this direction is posterior consistency. In a regression setting, studies of posterior consistency typically require that as the sample size goes to infinity, the random conditional densities are "close" to the data-generating conditional densities, almost surely with respect to the data-generating *joint* density. Of course this requires one to define a measure of closeness for the conditional densities, which is not

straightforward and is often measured by integrating classic measures of distance between density functions with respect to the data-generating marginal of x. Recent literature confirms properties of posterior consistency for some specific models (Hannah et al. [2011], Pati et al. [2012], Norets and Pelenis [2012b]). A subject of ongoing research is to verify these properties for the models developed in this thesis that improve predictive performance.

However, Bayesian nonparametric mixture models for regression, including the ones developed here, are highly flexible and likely to be consistent. Thus, while consistency properties provide important model validation, they are unlikely to be helpful in terms of model comparison, which further highlights the question of how to formally compare Bayesian nonparametric models. Stronger frequentist properties, such as convergence rates, could provide a solution, but there is currently no literature on this subject for the flexible Bayesian nonparametric regression mixture models that are studied here.

Instead, we aim to provide formal model comparison through predictive performance by formalizing our findings on prediction for the models of interest. For example, in Chapter 4, we discussed how the proposed EDP mixture model can be more efficient in exploiting the information present in the data, leading to smaller predictive estimation errors and tighter credible intervals. In this case, we aim to to quantify this gain in efficiency, under certain assumptions of the data-generating conditional densities.

The literature on predictive model comparison (San Martini and Spezzaferri [1984], Laud and Ibrahim [1995], Gelfand and Ghosh [1998]) is a starting point for our analysis. In addition, we intend to explore predictive properties, such as finite sample bounds on the probability that regression function or conditional density at some new covariate value is "close" to the truth. Ideally, these results would depend not only on the model but also on the hyperparameters and various aspects of the data including the sample size, dimension of the covariates, response type, and covariate types. Such results would greatly aid in the selection of the appropriate model and hyperparameters for the dataset at hand.

The models developed in this thesis were used to study the structure of tissue loss in Alzheimer's disease. In Chapter 5, we considered the diagnosis of AD based on the asymmetry of hippocampal volume and found evidence for both a "left-less-than-right" and a "right-less-than-left" pattern of atrophy. In Chapter 4, AD was diagnosed based on the volume and cortical thickness of several brain structures. The results were comparable, if not slightly better, than standard nonparametric regression techniques. This is an encouraging result that suggests that the EDP mixture model may be a useful extension of the flexible class of Bayesian nonparametric mixture models in high dimensions. In Chapter 6, we explored the dynamics of hippocampal volume as a function of age, sex, and disease status, and the results of our model confirmed the hypothesized sigmoidal behavior of hippocampal volume as a function of age.

In further studies, we intend to explore the diagnosis of AD based on a finer summary of the neuroimage, or possibly the entire neuroimage, and combine this with data obtained from other neuroimaging techniques and clinical and biological information. Another important study will involve investigating the dynamics of several AD biomarkers jointly. An initial study is under way to explore the dynamics of several well studied biomarkers during the early stages of AD, with the goal of determining the best biomarkers to use as outcome measures in clinical trials during early stages of AD. This is joint work with Anna Caroli and others from the Laboratory of Epidemiology and Neuroimaging, IRCCS San Giovanni di Dio-FBF, in Brescia, Italy.

Bayesian nonparametric mixture models for regression seem appropriate for these studies because of their flexibility and ability to capture the complex interactions terms that are likely to be present in the data. Any model properties that suggest improved predictive performance for a specific model in these applications would be very useful. Furthermore, neuroimaging datasets are extremely high-dimensional, and more so, as data from multiple imaging techniques are considered. Thus, a study of model properties for large p would be very interesting, and any future work that combines the flexibility of Bayesian nonparametric mixture models

with dimension reduction techniques would be useful.

Bibliography

R.P. Adams, I. Murray, and D.J.C. MacKay. Nonparametric Bayesian density modeling with Gaussian processes. 2009. URL `http://arxiv.org/abs/0912.4896`.

Alzheimer's Disease Education & Referral Center ADEAR. Alzheimer's disease fact sheet. *NIH Publication*, 11-6423, 2011.

C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2:1152–1174, 1974.

I. Antoniano Villalobos, S.K. Wade, and S.G. Walker. A Nonparametric Regression Model for the study of hippocampal atrophy in Alzheimer's disease. *Submitted*, 2012.

A.F. Barrientos, A. Jara, and F.A. Qunitana. On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Analysis*, 7:277–310, 2012.

A. Bhattacharya, G. Page, and D.B. Dunson. Density estimation and classification via Bayesian nonparametric learning of affine subspaces. *Journal of the American Statistical Association*, Revision submitted, 2012. Available at `http://arxiv.org/abs/1105.5737`.

D. Blackwell and J.B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

D.M. Blei and P.I. Frazier. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 2011.

L. Breiman, J.H. Friedman, R. Olshen, and C.J. Stone. *Classification and regression trees*. Wadsworth, Belmont, CA, 1984.

R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H.M. Arrighi. Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 3:186–191, 2007.

A. Caroli and G.B. Frisoni. Quantitative evaluation of Alzheimer's disease. *Expert Review of Medical Devices*, 6:569–588, 2009.

A. Caroli and G.B. Frisoni. The dynamics of Alzheimer's disease biomarkers in the Alzheimer's Disease Neuroimaging Initiative cohort. *Neurobiology of Aging*, 31:1263–1274, 2010.

H.A. Chipman, E.I. George, and R.E. McCulloch. BART: Bayesian additive regression trees. *Annals of Statistics*, 4:266–298, 2010.

Y. Chung and D.B. Dunson. Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104:1646–1660, 2009.

Y. Chung and D.B. Dunson. The local Dirichlet process. *Annals of the Institute for Statistical Mathematics*, 63:59–80, 2011.

D.M. Cifarelli and E. Regazzini. Problemi statistici nonparametrici in condizioni di scambiabilità parziale e impiego di medie associative. *Quaderni Istituto di Matematica Finanziaria, Università di Torino*, 12:1–36, 1978. English translation available at `www.unibocconi.it/wps/allegatiCTP/CR-Scamb-parz[1].20080528.135739.pdf`.

D.M. Cifarelli and E. Regazzini. De Finetti's contribution to probability and statistics. *Statistical Science*, 11:253–282, 1996.

D.M. Cifarelli, P. Muliere, and M. Scarsini. Il modello lineare nell'approccio Bayesiano non parametrico. *Istituto Matematico G. Castelnuovo, Universitá degli Studi di Roma La Sapienza*, 15, 1981.

B. Clarke, E. Fokoué, and H.H. Zhang. *Principles and theory for data mining and machine learning.* Springer Series in Statistics, New York, 2009.

R.J. Connor and J.E. Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64:194–206, 1969.

G. Consonni and P. Veronese. Conditionally reducible natural exponential families and enriched conjugate priors. *Scandinavian Journal of Statistics*, 28:377–406, 2001.

D.B. Dahl. Distance-based probability distribution for set partitions with applications to Bayesian nonparametrics. In *JSM Proceedings. Section on Bayesian Statistical Science*, Alexandria, VA, 2008. American Statistical Association.

P. Damien and S.G. Walker. Sampling from truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, 10:296–215, 2001.

C. Davatzikos, Y. Fan, X. Wu, D. Shen, and S.M. Resnick. Detection of prodromal Alzheimer's disease via pattern classification of MRI. *Neurobiology of Ageing*, 29:514–523, 2008a.

C. Davatzikos, S.M. Resnick, X. Wu, P. Parmpi, and C.M. Clark. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *Neuroimage*, 41:1220–1227, 2008b.

M. De Iorio, P. Müller, G.L. Rosner, and S.N. MacEachern. An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, 99:2205–215, 2004.

M. De Iorio, W.O. Johnson, P. Müller, and G.L. Rosner. Bayesian nonparametric non-proportional hazards survival modelling. *Biometrics*, 65:762–771, 2009.

R. De La Cruz, F.A. Quintana, and P. Müller. Semiparametric Bayesian classification with longitudinal markers. *Journal of the Royal Statistical Society, Series C*, 56:119–137, 2007.

D.G.T. Denison, C.C. Holmes, B.K. Mallick, and A.F.M Smith. *Bayesian methods for nonlinear classification and regression*. John Wiley & Sons, 2002.

P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *Annals of Statistics*, 7:269–281, 1979.

I. DiMatteo, D.R. Genovese, and R.E. Kass. Bayesian curve fitting with free-knot splines. *Biometrika*, 88:1055–1071, 2001.

K.A. Doksum. Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability*, 2:183–201, 1974.

J.A. Duan, M. Guindani, and A.E. Gelfand. Generalized spatial Dirichlet processes. *Biometrika*, 94:809–ï¿½825, 2007.

D.B. Dunson. Nonparametric Bayes local partition models for random effects. *Biometrika*, 96:249–262, 2009.

D.B. Dunson. Nonparametric Bayes applications to biostatistics. In N.L. Hjort, C. Holmes, P. Müller, and S.G. Walker, editors, *Bayesian nonparametrics*. Cambridge University Press, 2010.

D.B. Dunson and A.H. Herring. Semiparametric Bayesian latent trajectory models. *Technical Report, ISDS Discussion Paper 16, Duke University*, 2006.

D.B. Dunson and J.H. Park. Kernel stick-breaking processes. *Biometrika*, 95:307–323, 2008.

D.B. Dunson, N. Pillai, and J.H. Park. Bayesian density regression. *Journal of Royal Statistical Society, Series B*, 69:163–183, 2007.

D.B. Dunson, J. Xue, and L. Carin. The matrix stick breaking process: Flexible Bayes meta analysis. *Journal of the American Statistical Society*, 103:317–327, 2008.

D.B. Dunson, S. Petrone, and L. Trippa. Partially hierarchical Dirichlet mixtures for flexible clustering and regression. *Unpublished manuscript*, 2011.

M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90: 577–588, 1995.

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

J.H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19:1–67, 1991.

J.H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

G.B. Frisoni, N.C. Fox, C.R. Jr Jack, P. Scheltens, and P.M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6:67–77, 2010.

R. Fuentes-Garcia, R.H. Mena, and S.G. Walker. A probability for classification based on the mixture of Dirichlet process model. *Journal of Classification*, 27:389–403, 2010.

D. Geiger and D. Heckerman. A characterization of the Dirichlet distribution through global and local parameter independence. *Annals of Statistics*, 25:1344–1369, 1997.

A.E. Gelfand and S. Ghosh. Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85:1–13, 1998.

A.E. Gelfand, A. Kottas, and S.N. MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, pages 1021–1035, 2005.

J. Geweke and M. Keane. Smoothly mixing regressions. *Journal of Econometrics*, 138:252–290, 2007.

S. Ghosal and A.W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29:1233–1263, 2001.

S. Ghosal and A.W. van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35:1556–1593, 2007.

S. Ghosal, J.K. Ghosh, and R.V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27: 143–158, 1999.

J.K. Ghosh and R.V. Ramamoorthi. *Bayesian Nonparametrics*. Springer-Verlag, Springer Series in Statistics, New York, 2003.

S.J. Godsill. On the relationship between Markov chain Monte Carlo Methods for model uncertainty. *Journal of Computational and Graphical Statistics*, 10(2):230–248, 2001.

P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

J.E. Griffin and M. Steele. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 10:179–194, 2006.

J.E. Griffin and M. Steele. Bayesian nonparametric modelling with the Dirichlet process regression smoother. *Statistica Sinica*, 20:1507–1527, 2010.

J.E. Griffin, M. Kolossiatis, and M. Steele. Comparing distributions using dependent normalized random measure mixtures. *Working paper*, 2011.

L. Hannah, D. Blei, and W. Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12:1923–1953, 2011.

T.J. Hastie and R.J. Tibshirani. *Generalized additive models*. Chapman & Hall, London, 1 edition, 1990.

Orellana Y. Iglesias, P.L. and F.A. Quintana. Nonparametric Bayesian modelling using skewed Dirichlet processes. *Journal of Statistical Planning and Inference*, 139:1203–1214, 2009.

H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173, 2001.

C.R. Jr Jack, D.S. Knopman, W.J. Jagust, L.M. Shaw, Aisen P.S., M.W. Weiner, R.C. Petersen, and J.Q. Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurology*, 9:119–128, 2010.

C.R. Jr Jack, P. Vemuri, H.J. Wiste, S.D. Weigand, T.G. Lesnick, V. Lowe, K. Kantarci, M.A. Bernstein, M.L. Senjem, J.L. Gunter, B.F. Boeve, J.Q. Trojanowski, L.M. Shaw, P.S. Aisen, M.W. Weiner, R.C. Petersen, and D.S. Knopman. Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. *Archives of Neurology*, 69:856–867, 2012.

R.A. Jacobs and M.I. Jordan. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

R.A. Jacobs, M.I. Jordan, S. Nowlan, and G.E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:1–12, 1991.

A. Jara, E. Lesaffre, M. De Iorio, and F.A. Quintana. Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Annals of Applied Statistics*, 4:2126–2149, 2010.

R. Jenkins, N.C Fox, A.M. Rossor, R.J. Harvey, and M.N. Rossor. Intracranial volume and Alzheimer disease: Evidence against the cerebral reserve hypothesis. *Archives of Neurology*, 57:220–224, 2005.

M. Kalli, J.E. Griffin, and S.G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21:93–105, 2011.

C. Kang and S. Ghosal. Clusterwise regression using Dirichlet process mixtures. *Advances in Multivariate Statistical Methods*, pages 305–325, 2009.

S. Kloppel, C.M. Stonnington, C. Chu, B. Draganski, R.I. Scahill, J.D. Rohrer, N.C. Fox, C.R. Jr. Jack, J. Ashburner, and R.S.J. Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131: 681–689, 2008.

D.S. Knopman, S.T. DeKosky, J.L. Cummings, H. Chui, J. Corey-Bloom, N. Relkin, G.W. Small, B. Miller, and J.C. Stevens. Practice parameter: Diagnosis of dementia (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, 56:1143–1153, 2001.

M. Kolossiatis, J.E. Griffin, and M. Steele. On bayesian nonparametric modelling of two correlated distributions. *Statistics and Computing*, pages 1–15, 2011.

M.P. Laakso, H. Soininen, K. Partanen, M. Lehtovirta, M. Hallikainen, T. Hanninen, E.L. Helkala, P. Vainio, and P.J. Sr. Riekkinen. MRI of the hippocampus in Alzheimer's disease: sensitivity, specificity, and analysis of the incorrectly classified subjects. *Neurobiology of Ageing*, 19:23–31, 1998.

P.W. Laud and J.G. Ibrahim. Predictive model selection. *Journal of the Royal Statistical Society, Series B*, 57:247–262, 1995.

J.P. Lerch, J.C. Pruessner, A. Zijdenbos, H. Hampel, S.J. Teipel, and A.C. Evans. Focal decline of cortical thickness in Alzheimer's disease identifed by computational neuroanatomy. *Cerebral Cortex*, 15:995–1001, 2005.

A. Lijoi, B. Nipoti, and I. Prünster. Bayesian inference with dependent normalized completely random measures. *Technical Report, Collegio Carlo Alberto*, 2011.

A.Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, 12:351–357, 1984.

S.N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics - Simulation and Computation*, 23:727–741, 1994.

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

S.N. MacEachern. Dependent Dirichlet processes. *Technical Report, Department of Statistics, Ohio State University*, 2000.

S.N. MacEachern. Decision theoretic aspects of dependent nonparametric processes. In E. George, editor, *Bayesian Methods With Applications to Science, Policy, and Official Statistics*, pages 551–560. ISBA, 2001.

G.J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley series in probability and statistics: Applied probability and statistics. Wiley, 2000.

A. Mira and S. Petrone. Bayesian hierarchical nonparametric inference for change-point problems. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 5*. Oxford Univeristy Press, 1996.

J. Møller, A.N. Pettitt, R. Reeves, and K.K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

P. Muliere and S. Petrone. A Bayesian predictive approach to sequential search for an optimal dose: parametric and nonparametric models. *Journal of Italian Statistical Society*, 2:349–364, 1993.

P. Müller and F.A. Quintana. Nonparametric Bayesian data analysis. *Statistical Science*, 19:95–110, 2004.

P. Müller and F.A. Quintana. Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140:2801–2808, 2010.

P. Müller, A. Erkanli, and M. West. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 88:67–79, 1996.

P. Müller, F.A. Quintana, and G. Rosner. A method for combining inference across related nonparametric Bayesian models. *Journal of Royal Statistical Society, Series B*, 64:735–749, 2004.

P. Müller, G. L. Rosner, M. De Iorio, and S.N. MacEachern. A nonparametric Bayesian model for inference in related longitudinal studies. *Journal of the Royal Statistical Society, Series C*, 54:611–626, 2005.

P. Müller, F.A. Quintana, and A.L. Papoila. Cluster-specific variable selection for product partition models. *Submitted working paper*, 2012.

I. Murray, Z. Ghahramani, and D.J.C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press, 2006.

R.M. Neal. *Bayesian learning for neural networks*. Lecture Notes in Statistics. Springer, 1996.

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistcs*, 9:249–265, 2000.

A. Norets. Approximation of conditional densities by smooth mixtures of regressions. *Annals of Statistics*, 38:1733–1766, 2010.

A. Norets and J. Pelenis. Bayesian modeling of joint and conditional distributions. *Journal of Econometrics*, 168:332–346, 2012a.

A. Norets and J. Pelenis. Posterior consistency in conditional distribution estimation by covariate dependent mixtures. *revision requested by Econometric Theory*, 2012b. Available at `http://www.princeton.edu/ anorets/consmixreg.pdf`.

O. Papaspiliopoulos and G.O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95 (1):169–186, 2008.

J.H. Park and D.B. Dunson. Bayesian generalized product partition model. *Statistica Sinica*, 20:1203–1226, 2010.

D. Pati, D.B. Dunson, and S. Tokdar. Posterior consistency in conditional distribution estimation. *Submitted to the Annals of Statistics*, 2012. Available at `ftp://152.3.22.8/pub/WorkingPapers/10-17.pdf`.

S. Petrone and A.E. Raftery. A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statistics and Probability Letters*, 36:69–83, 1997.

S. Petrone and L. Trippa. Bayesian modeling via nested random partitions. In *Proceedings of the International Conference on Complex data modeling and computationally intensive statistical methods, September 14-16, 2009*, Milan, Italy, 2009. Politecnico di Milano.

S. Petrone, M. Guindani, and A.E. Gelfand. Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society, Series B*, 71:755–782, 2009.

A.N. Pettitt, N. Friel, and R. Reeves. Efficient calculation of the normalizing constant of the autologistic and related models on the cylinder and lattice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):235–246, 2003.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102:145–158, 1995.

J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900, 1997.

F.A. Quintana. A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*, 136:2407–2429, 2006.

F.A. Quintana. Linear regression with a dependent skewed Dirichlet process. 2011.

S. Rabe-Hesketh and A. Skrondal. *Multilevel and longitudinal modeling using Stata*. Stata Press, College Station, Texas, 2005.

R.V. Ramamoorthi and L. Sangalli. On a characterization of Dirichlet distribution. In S. Upadhyay, U. Singh, and D. Dey, editors, *Proceedings of the International Conference on Bayesian Statistics and its Applications, Jan. 6-8, 2005*, pages 385–397, Varanasi, India, 2006. Banaras Hindu University.

C.E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, Cambridge, MA, 2002. the MIT Press.

C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*. the MIT Press, 2006.

B.J. Reich and M. Fuentes. A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Annals of Applied Statistics*, 1:249–264, 2007.

L. Ren, L. Du, D.B. Dunson, and L. Carin. The logistic stick-breaking process. *Journal of Machine Learning and Research*, 12:203–239, 2011.

A. Rodriguez and D.B. Dunson. Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6:145–178, 2011.

A. Rodriguez and E. Horst. Bayesian dynamic density estimation. *Bayesian Analysis*, 3:339–366, 2008.

M.R. Sabuncu, R.S. Desikan, J. Sepulcre, B.T.T. Yeo, H. Liu, N.J. Schmansky, M. Reuter, M.W. Weiner, R.L. Buckner, R.A. Sperling, and B. Fischl. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Archives of Neurology*, 68:1040–1048, 2011.

A. San Martini and F. Spezzaferri. A predictive model selection criterion. *Journal of Royal Statistical Society, Series B*, 46:296–303, 1984.

D.W. Scott. *Multivariate density estimation: Theory, Practice, and Visualization*. John Wiley & Sons, Inc., Hoboken, NJ, 1992.

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

B. Shahbaba and R.M. Neal. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10:1829–1850, 2009.

F. Shi, B. Lui, Y. Zhou, C. Yu, and T. Jiang. Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of MRI studies. *Hippocampus*, 19:1055–1064, 2009.

M.D. Springer and W.E. Thompson. The distribution of products of beta, gamma and gaussian random variables. *Journal on Applied Mathematics*, 18:721–737, 1970.

M.A. Taddy and A. Kottas. A Bayesian nonparametric approach to inference for quantile regression. *Journal of Business and Economic Statistics*, 28:357–369, 2010.

Y.W. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet process. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

S.T. Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, 68:90–110, 2006.

S.T. Tokdar. Adaptive convergence rates of a Dirichlet process mixture of multivariate normals. 2011.

B. Vidakovic. *Statistical modelling by wavelets.* John Wiley & Sons, 2009.

S.K. Wade, S. Mongelluzzo, and S. Petrone. An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis*, 6:359–386, 2011.

S.K. Wade, S.G Walker, and S. Petrone. A predictive study of Dirichlet process mixture models for curve fitting. *Submitted*, 2012.

G. Wahba. *Spline models for observational data.* SIAM: Society for Industrial and Applied Mathematics, 1990.

S.G. Walker and P. Muliere. A bivariate Dirichlet process. *Statistics and Probability Letters*, 64:1–7, 2003.

S.G. Walker, A. Lijoi, and I. Prünster. On rates of convergence for posterior distributions in infinite-dimensional models. *Annals of Statistics*, 35:738–746, 2007.

M. West, P. Müller, and M. D. Escobar. Hierarchical priors and mixture models, with applications in regression and density estimation. *Aspects of Uncertainty: A Tribute to D.V. Lindley*, pages 363–386, 1994.

H. Wolf, M. Grunwald, F. Kruggel, S.G. Riedel-Heller, S. Angerho, A. Hojjatoleslami, A. Hensel, T. Arendt, and H.J. Gertz. Hippocampal volume discriminates between normal cognition; questionable and mild dementia in the elderly. *Neurobiology of Ageing*, 22:177–186, 2001.

Y. Wu and S. Ghosal. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, 2:298–331, 2008.

Y. Wu and S. Ghosal. The L_1-consistency of Dirichlet mixtures in multivariate density estimation. *Journal of Multivariate Analysis*, 101:2411–2419, 2010.