UNIVERSIDAD CARLOS III DE MADRID

Departamento de Teoría de la Señal y Comunicaciones

DOCTORAL THESIS

BAYESIAN NONPARAMETRIC MODELS FOR DATA EXPLORATION

Author: MÉLANIE NATIVIDAD FERNÁNDEZ PRADIER

Supervised by: FERNANDO PÉREZ CRUZ

July 2017

Tesis Doctoral:	BAYESIAN NONPARAMETRIC MODELS FOR DATA EXPLORATION
Autor:	Mélanie Natividad Fernández Pradier
Director:	Fernando Pérez Cruz
Fecha:	September 15th, 2017

Tribunal

Presidente: Joaquín Míguez Arenas

Vocal: Cédric Archambeau

Secretario: Daniel Hernández Lobato

Calificación:

“A person with a new idea is a crank until the idea succeeds.”

Mark Twain

“自分を信じる力、それが運命を変える力となる。”

不知火ゲンマ

Acknowledgements

A mis padres, Fernando y Marie-Thérèse, por su apoyo incondicional.
A Nii-san, por mostrarme el camino y forjarme el carácter.
A Javier, por estar siempre a mi lado.

When I was little, I remember trying to naïvely decipher my elder brother's notebook, which was full of integrals, and ending up crying because I was not able to understand such mysteries. Back then, I felt an immense wall raising in front of me, which I could finally demolish a few years later. Since then, I have vitally wandered, challenge after challenge, always seeking to confront the most fulfilling trials.[1]. Pick and shovel on my shoulder, I decided to leave the industry career path and start my new academic quest. Fall, winter, spring, and next was summer. The cherry trees blossomed late May, and this thesis needed to come to an end: I am ready to tear down this new facade.

But now more than ever, I am bitterly aware of the immensity of ignorance: the many things I do not know and that I will probably never get to know. Also now more than ever, I am able to comprehend the beauty of life and knowledge, and the exceptional talent of people around me, which have been an infinite (nonparametric) source of inspiration during these years. This thesis is not a masterpiece, but I have made endless sacrifices in the learning path, and put an immeasurable effort in its accomplishment, which is the best way I can now truly thank all those people who believed in me and stood by me all along.

To begin with, I would like to express my deep gratitude to my supervisor Fernando Pérez-Cruz for never losing faith in me, always granting me the freedom to pursue my own research direction, and teaching me the Yamamoto's challenging way of "working with delight". The signal theory and communications group at Universidad Carlos III de Madrid (UC3M) has been a fantastic learning environment where I have had the privilege to meet remarkable people. In particular, I would like to thank my academic siblings: Francisco J. R. Ruiz, for nurturing my passion for science, games and desserts; Isabel Valera, for her warm hugs and tenacity within research; Pablo M. Olmos, for his valuable mentorship in science, basketball and arcade games. Thanks to Victor Elvira for a shared complicity and feverish passion for multidisciplinary subjects; thanks to Pablo G. Moreno

[1] A certain pursuit for "suffering" is apparently a common pattern among PhD students.

for accepting me as an anonymous Bayesian despite my journeys into the frequentist domain. Also thanks to Jesus Fernandez-Bes, for being the living encyclopedia of the group (including the content of all BOEs since 1989). Many thanks to Alfredo Nazabal for his enduring peace of mind, all the enjoyable "procastinating" afternoons of research, Japanese/Chinese sessions, and go. Thanks to my dear labmates: Yanfang Liu, for enlightening my soul with craziness and creativity; O. Deniz Akyıldız, for increasing my thirst of knowledge and love for Leganés even further; Alejandro Lancho for stimulating me to run more and faster.

I am grateful to Tobias Koch, Gonzalo Vázquez, and Joaquín Míguez for their inspiring attitude towards science and constant good-nature. I would also like to thank Antonio Artés for adopting me in the GTS family, and for our open discussions about different ways of living and types of wines. Also thanks to the new generation, Pablo Bonilla and Pablo Moreno, for your sparking enthusiasm and steady willingness to make your utmost. Many thanks to Grace Villacrés, Francisco Hernando, David Ramírez, Gonzalo Ríos, and Alberto Valera for sharing an uncountable number of coffees and good chats together. Thank you to Paloma Vaquero for her sunny friendship and unshakable smile no matter what the difficulties might be. I would also like to thank Harold Molina-Bulla and Ana Hernando for their constant help along these years, as well as Marcelino Lázaro, for his guidance in the art of teaching.

Outside the UC3M, I feel very thankful to the GAPS research group within the signal processing department at the Universidad Politécnica de Madrid (UPM), which accepted me wholeheartedly in their scientific discussions with regularity. Thank you to Santiago Zazo for all the philosophical wanderings while running moderately across fields. Truly thanks to Sergio Valcarcel, whose positivity and kindness are beyond any limit; Marta Martínez, for her valuable friendship and contagious love for music; Pavle Belanovic, for showing me an endless enthousiasm for scientific and personal undertaking.

Special thanks to the European Marie Curie Initial Training Network sponsorship and colleagues within the network, with special mention to Felipe Llinarez, Dean Bodenham, Laetitia Papaxanthos, Cristóbal Esteban, Yi Zhong, Ramouna Fouladi, and Yunlong Jiao. I feel indebted for your friendship and enriching discussions via Skype, summer schools, research stays and conferences. I am also extremely grateful to my external collaborators: Francesca Milletti and Oscar Puig from Roche Diagnostics in New York; Viktor Stojkoski, Zoran Utkovski and Ljupco Kocarev from the Macedonian Institute of Science and Arts; Stephanie Hyland, Stefan Stark, Julia E. Vogt and Gunnar Rätsch from the Memorial Sloan-Kettering Cancer Center in New York; Kridsadakorn Chaichoompu, Fentaw Abegaz, and Kristel Van Steen from the University of Liège.

On the industry side, I would like to thank my former supervisor Thomas Kemp, who put his trust on me and gave me plenty of opportunities to grow internally. I also feel very grateful towards Naoki Kamimaeda and Masanori Miyahara for their incalculable guidance and support during my research internship in Japan, and towards all members of the former IAV department for receiving me with open arms in their group during a period of time that left a profound imprint on my spirit, strengthening my character and perseverance skills.

On the personal side, I would like to thank my dear friend Pedro A. De Mattos, whose antithetical vision of the world has enriched me enormously and led me to valuable compromises within society; also thanks to Luis Jou for displaying a regular, undisturbed passion for life and work. Thank you to Constanza Blanco for trying to make from the world a better place together. I am grateful to all the friends I met in Germany, specially in the kitchen from Allmandring 20C in Stuttgart (Timotheus, Tomoko, Simon, Haitham, Jonas, Diego, Chisato, among others),as well as my childhood friends who always believed in me (Sara, Raquel, Clara, Nuria and Alba). I cannot forget mentioning Guillermo, Berta, Andrés, and Elena, for the intense debates and wonderful trips together.

The achievement of this doctoral thesis would have never been possible without the unwavering support of my family and beloved ones. Thank you to my parents and brother, I owe you more than I could possibly express with words. Thanks to Pilar, Clara, and Tanis, for all their affection and understanding during these years. I would also like to thank all the little fur balls (rabbits, cats and dog) which have contributed with their expertise on stress management: Yan, Sid and Leia, Tigre and Nala, Gea, Puchi and Take, Taiko and the little Beibei. Finally, I am truly thankful to Javier Zazo for all the thoughtful discussions about physics, life, the universe, and beyond: you started a *river of savvy words in the blazing night drift*, and have been there for me since then, understanding me, encouraging me, and suffering me during the whole PhD.

All in all, thanks to all friends, family and colleagues that have been with me at some point in my life and that - I hope - will stay connected. This doctoral thesis is hopefully just the starting door to more exciting research and inspiring encounters in the future. Hopefully, ten years from now when I look into these pages, cracking down the PhD wall might appear as a child's play.

Abstract

Making sense out of data is one of the biggest challenges of our time. With the emergence of technologies such as the Internet, sensor networks or deep genome sequencing, a true *data explosion* has been unleashed that affects all fields of science and our everyday life. Recent breakthroughs, such as self-driven cars or champion-level Go player programs, have demonstrated the potential benefits from *exploiting* data, mostly in well-defined supervised tasks. However, we have barely started to actually *explore* and truly understand data.

In fact, data holds valuable information for answering most important questions for humanity: How does aging impact our physical capabilities? What are the underlying mechanisms of cancer? Which factors make countries wealthier than others? Most of these questions cannot be stated as well-defined supervised problems, and might benefit enormously from multidisciplinary research efforts involving easy-to-interpret models and rigorous data exploratory analyses. Efficient data exploration might lead to life-changing scientific discoveries, which can later be turned into a more impactful exploitation phase, to put forward more informed policy recommendations, decision-making systems, medical protocols or improved models for highly accurate predictions.

This thesis proposes tailored Bayesian nonparametric (BNP) models to solve specific data exploratory tasks across different scientific areas including sport sciences, cancer research, and economics. We resort to BNP approaches to facilitate the discovery of unexpected hidden patterns within data. BNP models place a prior distribution over an infinite-dimensional parameter space, which makes them particularly useful in probabilistic models where the number of hidden parameters is unknown a priori. Under this prior distribution, the posterior distribution of the hidden parameters given the data will assign high probability mass to those configurations that best explain the observations. Hence, inference over the hidden variables can be performed using standard Bayesian inference techniques, therefore avoiding expensive model selection steps.

This thesis is application-focused and highly multidisciplinary. More precisely, we propose an automatic grading system for sportive competitions to compare athletic performance regardless of age, gender and environmental aspects; we develop BNP models to perform genetic association and biomarker discovery in cancer research, either using genetic information and Electronic Health Records or clinical trial data; finally, we present a flexible infinite latent factor model of international trade data to understand the underlying economic structure of countries and their evolution over time.

Resumen

Uno de los principales desafíos de nuestro tiempo es encontrar sentido dentro de los datos. Con la aparición de tecnologías como Internet, redes de sensores, o métodos de secuenciación profunda del genoma, una verdadera *explosión digital* se ha visto desencadenada, afectando todos los campos científicos, así como nuestra vida diaria. Logros recientes como pueden ser los coches auto-dirigidos o programas que ganan a los seres humanes al milenario juego del Go, han demostrado con creces los posibles beneficios que podemos obtener de la *explotación de datos*, mayoritariamente en tareas supervisadas bien definidas. No obstante, apenas hemos empezado con la *exploración de datos* y su verdadero entendimiento.

En verdad, los datos encierran información muy valiosa para responder a muchas de las preguntas más importantes para la humanidad: ¿Cómo afecta el envejecimiento a nuestras aptitudes físicas? ¿Cuáles son los mecanismos subyacientes del cancer? ¿Qué factores explican la riqueza de ciertos países frente a otros? Si bien la mayoría de estas preguntas no pueden formularse como problemas supervisados bien definidos, éstas pueden ser abordadas mediante esfuerzos de investigación multidisciplinar que involucren modelos fáciles de interpretar y análisis exploratorios rigurosos. Explorar los datos de manera eficiente abre potencialmente la puerta a un sinnúmero de descubrimientos científicos en diversas áreas con impacto real en nuestras vidas, descubrimientos que a su vez pueden llevarnos a una mejor explotación de los datos, resultando en recomendaciones políticas adecuadas, sistemas precisos de toma de decisión, protocolos médicos optimizados o modelos con mejores capacidades predictivas.

Esta tesis propone modelos Bayesianos no-paramétricos (BNP) adecuados para la resolución específica de tareas exploratorias de los datos en diversos ámbitos científicos incluyendo ciencias del deporte, investigación contra el cáncer, o economía. Recurrimos a un planteamiento BNP para facilitar el descubrimiento de patrones ocultos inesperados subyacentes en los datos. Los modelos BNP definen una distribución a priori sobre un espacio de parámetros de dimensión infinita, lo cual los hace especialmente atractivos para enfoques probabilísticos donde el número de parámetros latentes es en principio desconocido. Bajo dicha distribución a priori, la distribución a posteriori de los parámetros ocultos dados los datos asignará mayor probabilidad a aquellas configuraciones que mejor explican las observaciones. De esta manera, la inferencia sobre el espacio de variables ocultas puede realizarse mediante técnicas estándar de inferencia Bayesiana, evitando el proceso de selección de modelos.

Esta tesis se centra en el ámbito de las aplicaciones, y es de naturaleza multidisciplinar. En concreto, proponemos un sistema de gradación automática para comparar el rendimiento deportivo de atletas independientemente de su edad o género, así como de otros factores del entorno. Desarrollamos modelos BNP para descubrir asociaciones genéticas y biomarcadores dentro de la investigación contra el cáncer, ya sea contrastando información genética con la historia clínica electrónica de los pacientes, o utilizando datos de ensayos clínicos; finalmente, presentamos un modelo flexible de factores latentes infinito para datos de comercio internacional, con el objetivo de entender la estructura económica de los distintos países y su correspondiente evolución a lo largo del tiempo.

Contents

List of Acronyms

3P-IBP	tree-parameter Indian buffet process.
3RBeP-PFA	three-parameter restricted Bernoulli process Poisson factor analysis.
ADCC	antibody-dependent cytotoxicity.
ADDP	atom-dependent Dirichlet process.
AGS	accelerated Gibbs sampling.
BeP	Bernoulli process.
BeP-PFA	Bernoulli process Poisson factor analysis.
BNP	Bayesian nonparametric.
BP	Beta process.
C-IBP	case-control Indian buffet process.
CC	case-control.
CRF	Chinese restaurant franchise.
CRM	completely random measure.
CRP	Chinese restaurant process.
dBeP-PFA	dynamic Bernoulli process Poisson factor analysis.
DDP	dependent Dirichlet process.
DNA	deoxyribonucleic acid.
DP	Dirichlet process.
EHR	electronic health record.

FDR	false discovery rate.
FFBS	forward-filtering backward-sampling.
FWER	family-wise error rate.
GDP	gross domestic product.
GLFM	general latent feature model.
GMM	Gaussian mixture model.
GP	Gaussian process.
GPC3	Glypican-3.
GWAS	genome-wide association study.
H-ADDP	hierarchical atom-dependent Dirichlet process.
H-PFA	hierarchical Poisson factor analysis.
HCC	hepatocellular carcinoma.
HDP	hierarchical Dirichlet process.
HS	harmonized system.
IBP	Indian buffet process.
IMoE	infinite mixture of experts.
IMoGGP	infinite mixture of global Gaussian processes.
LLH	log likelihood.
LMM	linear mixed model.
MCMC	Markov chain Monte Carlo.
mIBP	Markov Indian buffet process.
ML	machine learning.
MSE	mean square error.
NNMF	non-negative matrix factorization.
PFA	Poisson factor analysis.

PFS	progression free survival.
PGAS	particle Gibbs with ancestor sampling.
PGDS	Poisson Gamma dynamical system.
PMF	probabilistic matrix factorization.
R-IBP	restricted Indian buffet process.
RCA	revealed comparative advantage.
sBeP-PFA	sparse Bernoulli process Poisson factor analysis.
SITC	standard international trade classification.
SNP	single-nucleotide polymorphism.
sp-DDP	single-p dependent Dirichlet process.
SVD	singular value decomposition.
tGaP-PFA	thinned Gamma process Poisson factor analysis.
UMLS	unified medical language system.
WMA	world master of athletics.

1 Introduction

1.1 Motivation

We are living an exciting new era of science characterized by massive amounts of data. Every second, 2.9M emails are sent, 73 products are ordered on Amazon, and 20 minutes of video are uploaded to Youtube.[1] According to IBM, healthcare data double every 24 hours, from which around 80% is unstructured, waiting to be analyzed [166]. Recent advances in the field of deep learning have proved efficient in *exploiting* such huge amounts of data, bringing solutions to well-defined supervised tasks. Nonetheless, we are still far from getting the utmost out of data, specially in unsupervised scenarios, by *exploring* it and extracting valuable insights from reality so far unknown.

This is specially manifest in the field of medicine: although tons of data are available for each patient, most diagnoses and treatments still remain untailored to the needs of each individual. A much bigger gain might be expected if we manage to turn data into meaningful, **interpretable knowledge** first. This thesis contributes to this endeavor by focusing on probabilistic methods and inference algorithms for **data exploration**.

[1] Source: `http://www.globaldots.com/big-data-promise-hype-challenges/`

Three important obstacles need to be addressed in order to extract valuable information from data. First of all, *relationships between observations are typically highly complex*. Take for instance biological mechanisms causing a certain disease or sociological factors affecting in a presidential election. To analyze such data, we need flexible and easy-to-interpret models able to extract the essential information. Here, we focus on probabilistic graphical models, i.e., generative models which represent the input distribution of the data using latent variables. The use of latent variables allow us to abstract from reality and to understand statistical dependencies in an easy manner.[2] Graphical models are thus particularly suitable for collaboration across fields, as they allow for a bi-directional communication between domain experts and machine learning researchers, e.g., for model design and validation. These models have additionally the potential of delivering previously known information together with novel aspects of the data that were so far unknown. If a model is interpretable, that is, if a model can provide explanations or easy-to-understand information concerning its behavior, researchers will trust more its predictions [104, 175].

The second problem refers to having *small data within big data*. Observations never have the exact same contextual properties. In healthcare applications, disease evolution or drug effects strongly depend on individual characteristics, to the point that most major drugs are known to be effective in only 25 to 60 percent of patients, and more than 2 million cases of adverse drug reactions occur annually in the United States, including 100,000 deaths [218]. Small data also appear in the shape of outliers (e.g., patients suffering from rare diseases), or just due to missing observations. Furthermore, privacy and ethical concerns might difficult gathering huge amounts of data, e.g., we will never be able to run clinical trials of arbitrary sizes. Bayesian approaches address this issue through integration to compute posterior estimates, eliminate nuisance variables or missing data, and average models for prediction [11]. Within this framework, dependency structures can be incorporated to efficiently share information across varying observations.

The last challenge is that the number and complexity of *statistical hypotheses grow with data*, due to the so-called "curse of dimensionality" [18]. As a consequence, most exploratory studies in empirical sciences are not easy to replicate, once simulation conditions are slightly different. The growing number of statistical hypotheses might lead to a rising amount of false positives, e.g., artifact discoveries due to either chance or undesirable confounding factors. To address this problem, machine learning approaches should be able to generalize to unseen observations, without over-fitting. We may address this problem by incorporating tools from classical statistical methods, including assessment of statistical significance, multiple hypothesis testing or confounder correction [216].

[2] Synthetic samples can also be generated to validate the model and analyze its properties.

1.2 Scientific Aims and Perspective

This doctoral thesis deals with Bayesian nonparametric (BNP) models for data exploration. The Bayesian framework is particularly appealing in its ability to capture uncertainty of inferred parameters and avoid overfitting. Moreover, the nonparametric property refers to the ability of such models to automatically adapt their complexity depending on the amount of available data [61]. The number of latent variables is potentially unbounded, a priori unknown, and is also learned from the data. Another interesting aspect concerns the discrete nature of the random measures underlying BNP models, which make them sparse, and thus, naturally easy to interpret.

Ultimately, our objective is to develop BNP models and their corresponding inference algorithms to help experts in other fields get valuable insights from their data, in applications that have a strong impact on society. This research is application-driven, in the sense that we first understand the needs and challenges in a certain field and then try to give solutions through model abstractions. To reach that goal, we might need to improve existing models or design new ones, together with their respective inference algorithms. These resulting models might then generalize to other interesting applications which we did not consider in a first place. In short, we find inspiration for new models and inference approaches in real important problems for society.

Data exploratory purposes. Data exploration comes in different forms in the machine learning community: Principal component analysis and factor analysis are linear methods that provide non-sparse solutions with strong Gaussianity assumptions. Local-linear embedding [181], Isomap [205] and Gaussian process latent variable models [111] learn non-linear manifolds in high dimensional spaces with non-sparse features; non-negative matrix factorization [91] provides a low dimensional sparse representation of the data. Also, BNP models can be used for clustering [50] and sparse feature analysis [72], in which the underlying latent dimension is unknown. There are many applications that have benefited from data exploratory analyses, including market basket analysis [38], computer vision [220], genomics [27, 107, 28], social sciences [132], and psychiatry [182].

Although the flexibility of BNP models makes them particularly attractive for experts in other fields, obtaining *interpretable* results might be an even stronger requirement. In this thesis, we refer to *interpretability* as the "ability to explain or to present novel information in understandable terms to a human" [41]. Most BNP models are described as general priors [188, 125, 72] that might not give easy-to-interpret solutions if applied blindly, even if they provide accurate predictions. In order to additionally obtain interpretable results, e.g., meaningful structures for data exploration, we need to specify the priors and likelihood in a way that points towards the sought explanation by including

our prior knowledge and generative assumptions of the data in an adequate manner.

In this way, the first insights of the obtained results should not be foreign to us. This makes models trustworthy for experts in other fields that do not know about machine learning or statistics, so other conclusions that were not common knowledge can be taken as plausible. At this stage we are able to formulate hypotheses that can be tested with future data and can provide previously unknown insights about the given problem. Such procedure avoids the frequent black-box flavor found in other methods, facilitating collaboration across fields. Examples of such interdisciplinary efforts using BNP models can be found in psychiatry [182], genetics [222], biostatistics [46], computer vision [64], econometry [144] or musicology [174]. In this thesis, we are interested in practical data exploration applications of BNP models that specially benefit from their flexibility. This thesis puts special focus on model design and encoding of appropriate assumptions, which are crucial to bring an interpretable solution for each problem at hand.

1.3 Contributions

This thesis is multidisciplinary, bringing novel solutions to deal with real-world problems in the fields of sport sciences, cancer research and economics. Throughout this thesis, we address the following points:

(A) improve model interpretability via prior and likelihood design, e.g., imposing structure or sparsity in the solution space.

(B) increase model flexibility and ability to share information across samples through the implementation of dependent models.

(C) get replicable results by combining Bayesian approaches with classical statistical methods.

The contributions of this thesis have also been or will be partially published in [159, 157, 160, 154, 158, 210, 208, 161]. These correspond to extensions of existing BNP models across diverse research areas, with application to the problems of fairness in athletic competitions, biomarker and genetic association discovery in cancer research, and analysis of the economic structure of countries via their export portfolios.[3] We summarize our contributions below.

Fairness in Athletic Competitions

In order to study the impact of age, gender and environment on runner performance, we present a dependent infinite mixture model for density estimation of stratified data [159]. The novelty of

[3] Our contribution in [156] is out of scope of this thesis.

this work relies not only on the application, but also on the technical steps (non-trivial structural assumptions) to obtain interpretable results and share information across athletes. Our analysis delivers valuable information for sport science experts, as well as a fair system to compare runners, regardless of their age and gender, that could be directly incorporated in regular sport events. The presented methodology is general to compare group densities in applications having a certain evolutionary or competitive trait, such as in pediatrics (e.g., comparison of children population according to weight and height), social sciences (e.g., analysis of gender impact on actual salary income across countries), or epistemology (e.g., assessment of environment effect on scientific output). This work has additionally led to the development of computational structures using Hadoop and Spark to speed up inference [155],[4] and opened a new research line on non-linear regression problems [157], as described in Section 3.6.

Meaningful Discoveries in Cancer Research

Drug effect assessment through biomarker discovery in clinical trials. We propose a general BNP approach for biomarker discovery and subpopulation characterization in clinical trials. Our model has been used to help expert oncologists understand conditions for drug effectiveness. It also incorporates statistical techniques to account for false positives, and it separates drug effects from natural prognostic factors by sharing information among patients in a structured manner. We demonstrate the usefulness of our novel approach on a randomized phase II case-control study of a cutting-edge immunotherapy treatment against liver cancer, in collaboration with Roche Diagnostics. Not only did our method find already well-known statistically significant biomarkers, but it also discovered new ones that could not be found with previous approaches, opening the door to the development of a new drug for a subgroup of liver cancer patients [158]. The proposed model is an extension of the general latent feature model [210], for which we have also contributed with further empirical validation analyses [208] and a user-friendly C++ software release with wrappers for Matlab and Python (an R package is currently under development) on Github.[5]

Finding genetic associations with clinical features for enhanced diagnosis. In order to understand cancer mechanisms and their interactions with patients' phenotypes and environment, we analyze cancer-patients data from the Memorial Sloan-Kettering Cancer Center in New York. The database contains information from electronic health records and detailed genetic data obtained through deep genome sequencing. We look for associations between gene mutations and clinical

[4] This research line is out of scope of this thesis.

[5] Available at `https://github.com/ivaleraM/GLFM`.

features in cancer by first finding an appropriate nonparametric representation of the patient population via latent features, and then performing classical statistical tests on subsequent partitions of patients. The proposed BNP model allows for a joint analysis of multiple cancer types, and is able to deal with phenotype heterogeneity, small cohort size, epistatic and pleiotropic effects[6], as inspired by [154]. The proposed model is compared against classical alternatives, including linear mixed models where clinical features were tested for associations against the mutated genes [160].

Analysis of World Trade

Finally, we propose a BNP approach to analyze international trade. Our objective is to understand economic growth of countries, i.e., which factors make countries wealthier than others, and how these countries acquire such capabilities over time, with the ultimate goal of issuing economic policy recommendations. We first propose a flexible scheme for the static scenario, that incorporates relaxed prior assumptions on the activation of features in the latent space, in agreement with reality [161]. The relaxed prior allows each country to exhibit wider variations in the number of active features (reflecting rich vs poor countries), as well as more flexible *a priori* distributions in the global activation of features (accounting for simple vs specific capabilities). Second, we propose a dynamic extension to analyze the temporal evolution of countries' economies over time. We incorporate a Markovian structure over the features to account for time-varying feature activations for each country.

1.4 Organization

The remainder of this thesis is organized as follows. In Chapter 2 reviews the basics behind BNP models. In particular, we introduce the stochastic processes that will be used as building blocks along this thesis: the Dirichlet process (DP), and the Beta process (BP). The rest of the chapters are devoted to our contributions. Chapter 3 describes a dependent BNP model for marathon modeling. Chapter 4 and 5 develop BNP approaches for data exploratory analyses in the context of personalized medicine for cancer research. In Chapter 4, we focus on the clinical trials scenario, and bring a powerful tool to characterize subpopulations and identify valuable biomarkers to help expert oncologists in their research. In Chapter 5, we propose a joint hierarchical model for both clinical records and genetic data, in order to identify novel genetic associations with clinical features across different types of cancer. Chapter 6 develops static and dynamic Poisson factor analysis (PFA) schemes to understand the economic structure of countries via international trade. Finally, Chapter 7 is devoted to the conclusions and future lines of research.

[6]This work received a Spotlight Talk Award at the 9th Annual Machine Learning Symposium in New York, 2015.

2

Overview of Bayesian Nonparametrics

2.1 Introduction

Data in the real world typically involves some source of uncertainty. This uncertainty may come from noisy measurements, incomplete information, or from the finite size of datasets [62]. Probability theory has proven to be effective for understanding such data in terms of *degrees of belief*[1], establishing a consistent framework for quantification and manipulation of uncertainty. Models that incorporate random variables and probability distributions to quantify degrees of certainty are called *probabilistic Bayesian models*, and are at the foundation of pattern recognition. Such models constitute an important tool in all areas of science as a way to develop statistical algorithms for making predictions and learning hidden structures from data [18].

Bayesian approaches consider model parameters as unobserved random variables instead of deterministic values. The Bayesian paradigm allows to incorporate a priori knowledge of the world and desirable constraints over the solution space through the prior, as well as to account for uncertainty in the estimation of model parameters [11]. Within Bayesian models, *latent variable models* consist

[1] An alternative interpretation of probability is the frequentist point of view, where the probability of an event corresponds to the limit of its relative frequency in a large number of trials.

of hidden and observed variables, where the hidden variables encode underlying patterns in our data (typical applications include clustering, feature modeling, and topic modeling). Such models have had a major impact on numerous applied fields, including computational biology, economics, natural language processing, social network analysis, and computer vision [20].

Latent variable models rely on the *Bayes theorem*. Let $\mathbf{X}$ design the observed data, and $\mathbf{\Theta}$ be the set of unobserved variables. The Bayes theorem provides an elegant way to relate data-based evidence, encoded in the likelihood $p(\mathbf{X}|\mathbf{\Theta})$ with prior assumptions on the latent variables via the prior $p(\mathbf{\Theta})$. The Bayes theorem can be stated as:[2]

$$p(\mathbf{\Theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{\Theta})p(\mathbf{\Theta})}{p(\mathbf{X})}, \tag{2.1}$$

where $p(\mathbf{\Theta}|\mathbf{X})$ is the posterior distribution, i.e., the conditional distribution of the latent variables given the observed data, and $p(\mathbf{X})$ is the evidence or marginal distribution of the data under this model [18]. The posterior distribution can be used to explore, summarize, and form predictions about the data [20]. When the number of observations tends to infinity, the likelihood term $p(\mathbf{X}|\mathbf{\Theta})$ dominates over the prior.[3] In that respect, a prior can be understood as a "constraint" over the solution space whose influence might become weaker the more data we see.

This thesis focuses on latent variable models developed according to the "Box's loop" probabilistic pipeline [20], illustrated in Figure 2.1. This is an iterative loop process for data analysis, which consists of three fundamental stages: model formulation, inference, and model criticism. According to this scheme, we should first formalize our assumptions into a simple model to fit knowledge domain of the problem at hand, including hidden structure which we believe exists in the data. Second, we can use an inference algorithm to approximate[4] the posterior distribution and analyze the data under the assumptions encoded through the priors and likelihood. Finally, we should assess whether the analysis succeeds or fails, i.e., revise whether the model gives accurate predictions or insights that are consistent with current expert knowledge, and repeat the cycle if required.

Why nonparametrics? Most machine learning problems consist in learning an appropriate fixed set of parameters within a model class given the training data. Typically, practitioners fit several models with a different number of parameters, and use a separate validation set to determine the most adequate number of parameters. Determining appropriate model classes is referred to as *model*

[2]Note that the Bayes theorem is a general expression for any two random variables. We here further incorporate an asymmetric interpretation of the theorem, which lies at the foundation of Bayesian statistics.

[3]This does not mean that the posterior distribution will be consistent, i.e., that its mean will tend to the true values. Additional conditions would be necessary for this to hold.

[4]Computing the posterior analytically turns out to be intractable most of the time, so we need an approximation.

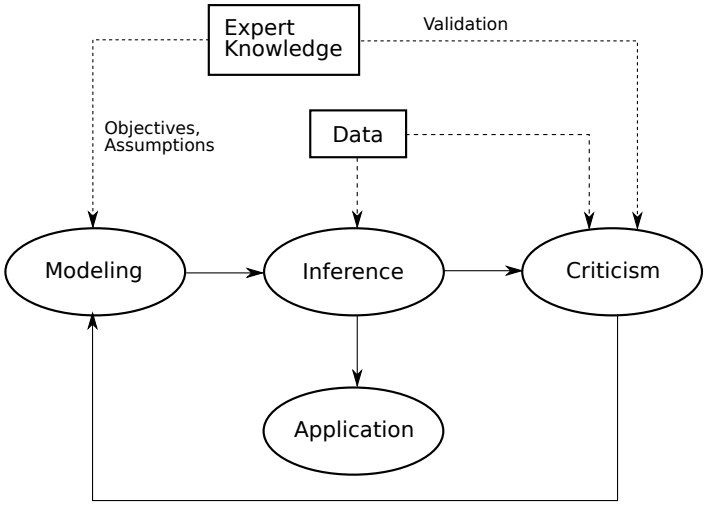


Figure 2.1: **Box's loop probabilistic pipeline for research.** First, we formalize the problem, and include appropriate assumptions based on expert knowledge. Second, we perform inference to learn the posterior distribution of the hidden variables. We can then use the posterior, i.e., apply the model, to make predictions and explore the data. Third and last, we revise the model using expert knowledge, accounting for weaknesses and strengths, eventually repeating the Box's loop if necessary.

selection in the literature [18]. Model selection is of fundamental concern for machine learning practitioners, not only for avoidance of over-fitting and under-fitting, but also for discovery of the appropriate causes and structures underlying data. Some examples of model selection and adaptation include: selecting the number of clusters in a clustering problem, the number of latent features in a matrix factorization problem, or the complexity (degrees of freedom) of a certain function in nonlinear regression.

Addressing model selection by fitting multiple models of varying complexity is computationally expensive. Nonparametric models constitute an alternative approach where model complexity is also learned from data. While parametric models are fully characterized by a fixed number of parameters, the number of parameters in nonparametric models grows with the amount of training data [135]. Nonparametric approaches are memory-based, in the sense that the amount of stored information into the model is proportional to the number of observations. For instance, the k-nearest neighbors method is nonparametric in the sense that it classifies the unseen instance based on the k points in the training set which are nearest to it (we need to store the whole training set for classification).[5]

Also, fitting a Gaussian mixture model (GMM) with a fixed number of Gaussians is a parametric approach for density estimation. A nonparametric version would be the Parzen window estimator, which centers a Gaussian at each observation, and hence uses one mean parameter per observa-

[5]Note that nonparametric refers to the nature of the method, it does not mean that the method has no parameters, e.g., the k-nearest neighbors has one parameter k.

tion [18]. Nonparametric methods have become popular in classical (non-Bayesian) statistics [215]. Another well-known example is the support vector machine [33], which has been widely applied to handle classification and regression problems. In model design, there is often a trade-off between flexibility and computational/storage costs. An example of such is the use of inducing points in the inference of Gaussian processes (GPs). Although a GP is a nonparametric prior over the space of functions, inducing points play the role of "local summaries" of the function and help limit the memory of the model, making it possible to scale inference to higher datasets [165].

2.2 Intuition behind Bayesian Nonparametric Models

Bayesian nonparametric (BNP) models made their debut in the 70s, with the seminal papers of Ferguson [53] and Antoniak [12] among others,[6]. Although mathematically elegant, inference was too problematic at that time, so BNP models did not pick up much attention in the community until the 90s, thanks to the introduction of Markov chain Monte Carlo (MCMC) methods [59]. Around ten years later, variational inference made its appearance in the field [102], and allowed BNP models to further scale to higher amounts of data, making them even more popular in the 2010s [21, 42]. Recent advances in the field of deep learning have taken the main interest of the machine learning community, putting BNP models in the background. Although powerful, the spread of BNP models is still moderate, for computational complexity still remains an important drawback. Several lines of research address this issue, both from the perspective of MCMC approaches (by including adaptive subsampling or stochastic gradient dynamics) [11] or scaling up variational inference techniques via noisy stochastic gradients [90].

BNP models combine the benefits of Bayesian methods with the flexibility of nonparametric approaches [61]. Instead of specifying a closed-form model, BNP techniques place probability mass on an infinite range of models and let the inference procedure select those configurations that best fit the data, in order to provide competitive predictions or density estimations [143]. Thus, BNP models provide a useful tool for problems in which the number of unknown hidden variables is itself unknown and can be learned using standard Bayesian inference techniques. Given our prior beliefs, the posterior in a parametric model is constrained to a closed family, whereas BNP models allow for more flexible shapes in the posterior distribution. Although a BNP model is characterized by an infinite-dimensional parameter space, only a finite subset of the available parameters is used for any given finite dataset. This subset will generally grow with the dataset size.

[6] De Finetti's theorem (which will be described later) dates from the 30s [35] but this was not fully exploited in the context of BNP approaches until decades later.

An illustrative example. In order to get a better intuition for BNP models, let us consider a finite GMM and its non-parametric Bayesian counterpart. For simplicity, we will assume unknown mixture means and known covariance matrix, e.g., $\mathbf{\Sigma} = \sigma_x \mathbf{I}$ where $\mathbf{I}$ refers to the identity matrix. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\top$ be the input matrix for N input observations $\mathbf{x}_n \in \mathbb{R}^{D\times 1}$, where $n = 1, \ldots, N$. A finite GMM of K clusters assumes the following likelihood distribution:

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N|\mathbf{\Phi}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \sigma_x \mathbf{I}), \qquad \mathbf{\Phi} = \{(\pi_k, \boldsymbol{\mu}_k)_{k=1}^K\}, \tag{2.2}$$

where $\mathbf{\Phi}$ refers to the set of hidden variables, $\boldsymbol{\mu}_k$ is the k-th cluster parameter (mean vector), and π_k is the mixture weight for cluster k. Given $\mathbf{X}$, our aim is to learn the distribution of $\mathbf{\Phi}$. Inferring the joint posterior distribution $p(\mathbf{\Phi}|\mathbf{X})$ directly is an intractable problem which cannot be solved analytically. To address such issue, a common practice is to introduce an auxiliary cluster assignment variable z_n for each observation $\mathbf{x}_n$ that indicates the mixture to which $\mathbf{x}_n$ belongs, e.g., $p(x_n|z_n) = \mathcal{N}(\boldsymbol{\mu}_{z_n}, \sigma_x \mathbf{I})$. Equation 2.2 can then be rewritten as:

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N|\mathbf{\Phi}) = \prod_{n=1}^{N} p(\mathbf{x}_n|z_n, \boldsymbol{\mu}_{1:K}), \qquad \mathbf{\Phi} = \{(\pi_k, \boldsymbol{\mu}_k)_{k=1}^K,\ (z_n)_{n=1}^N\}, \tag{2.3}$$

If we now assume a prior distribution over $\mathbf{\Phi}$, we can write the complete joint probability as

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \mathbf{\Phi}) = \Big(\prod_{n=1}^{N} p(\mathbf{x}_n|z_n, \boldsymbol{\mu}_{1:K}) p(z_n|\pi_k) \Big) \Big(\prod_{k=1}^{K} p(\boldsymbol{\mu}_k) \Big) p(\boldsymbol{\pi}), \tag{2.4}$$

and formulate a sequential generative process for $\mathbf{X}$ such that $p(\mathbf{X}|\mathbf{\Phi})$ is given by (2.2) or (2.3),

$$\boldsymbol{\pi} \sim \text{Dirichlet}(a_1, \ldots, a_K) \tag{2.5}$$

$$\forall k = 1, \ldots, K$$

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{\Sigma}_0) \tag{2.6}$$

$$\forall n = 1, \ldots, N$$

$$z_n \sim \text{Categorical}(\boldsymbol{\pi}) \tag{2.7}$$

$$\mathbf{x}_n|z_n \sim \mathcal{N}(\boldsymbol{\mu}_{z_n}, \sigma_x \mathbf{I}). \tag{2.8}$$

$\boldsymbol{\pi}$ refers to the vector of mixture weights $[\pi_1, \ldots, \pi_K]$, $\boldsymbol{\mu}_0$, $\mathbf{\Sigma}_0$ and $\boldsymbol{a} = \{a_1, \ldots, a_K\}$ are the model *hyperparameters*, e.g. prior parameters for the distribution of latent variables, and the Categorical

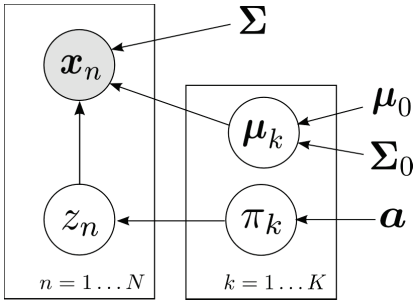


Figure 2.2: **Graphical representation for a finite GMM.** Nodes in a circle are random variables, grey nodes are observed ones, and white nodes are unobserved variables. Model (hyper)parameters are not circled. An infinite GMM has the same graphical representation but replacing $k = 1 \dots \infty$.

distribution is equivalent to a Multinomial distribution with number of draws equal to one. In this example, we have two types of latent variables: global variables (π_k and $\boldsymbol{\mu}_k$ for each k-th cluster) and local variables (one cluster assignment z_n for each observation $\mathbf{x}_n$). Figure 2.2 shows the corresponding graphical representation for a finite GMM with K clusters.

In contrast, a non-parametric GMM allows for a potentially unbounded number of clusters K, and can be obtained by using an infinite-dimensional distribution over the parameter space, i.e., by replacing (2.5) and (2.6) jointly by a stochastic process called the Dirichlet process (DP). The generative formulation in this case is given by:[7]

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\mu}_k} \sim \mathrm{DP}(\alpha, H) \tag{2.9}$$

$$\forall n = 1, \dots, N$$

$$\boldsymbol{\theta}_n \sim G \tag{2.10}$$

$$\mathbf{x}_n | \boldsymbol{\theta}_n \sim \mathcal{N}(\boldsymbol{\theta}_n, \sigma_x \mathbf{I}), \tag{2.11}$$

where $\boldsymbol{\theta}_n$ is the observation-specific cluster mean vector corresponding to observation n (using the previous cluster assignment variables, $\boldsymbol{\theta}_n = \boldsymbol{\mu}_{z_n}$), and G is a random measure corresponding to a realization of the DP which can be written as an infinite sum of *sticks* or *atoms* of weights π_k at locations $\boldsymbol{\mu}_k$, where $k = 1, \dots, \infty$ and $\sum_{k=1}^{\infty} \pi_k = 1$. A DP is a stochastic process parameterized by a concentration parameter α and base measure H (further details can be found in Section 2.3.1). Figure 2.3 shows a draw example $G \sim \mathrm{DP}(\alpha, H)$, where H defines a distribution for the atom locations, e.g., in our particular example $H \doteq \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, and α plays the role of an "inverse variance" parameter, such that small values of α result in fewer sticks with high weights, whereas bigger values of α "spread" the probability mass all over, resulting in a bigger number of sticks with

[7] Formal definitions will be provided in Section 2.3, we here provide an intuition for introducing BNP models.

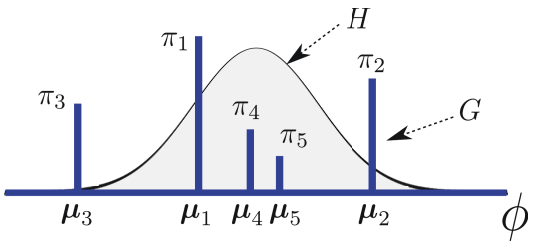


Figure 2.3: **A draw example G from a Dirichlet process.** We only depict the sticks with highest weights, although the number of sticks in infinite. Drawing G is equivalent to independently drawing each $\boldsymbol{\mu}_k \sim H$, and $\boldsymbol{\pi}$ from a certain *stick-breaking process*, which will be introduced in 2.3.1.

smaller weights. Given G, the generative process consists of sampling one local variable $\boldsymbol{\theta}_n$ for each data point $\mathbf{x}_n$. Since G is discrete with probability one, $\boldsymbol{\theta}_n$ will take repeated values in proportion to $\boldsymbol{\pi}$, resulting in different clusters of observations. Even if the number of atoms in G is infinite, a dataset is ultimately finite, and thus, the number of non-empty clusters (locations $\boldsymbol{\mu}_k$ for which $\exists j$ s.t. $\boldsymbol{\theta}_j = \boldsymbol{\mu}_k$) will be bounded. In a streaming application, sampling an observation-specific variable $\boldsymbol{\theta}^*$ for a new observation $\mathbf{x}^*$ could either return a cluster mean from which we have already sampled before, or a completely new cluster with some probability π_{new} that is proportional to α.

Generally speaking, the central idea behind BNPs is the replacement of classical finite-dimensional prior distributions with general stochastic processes, allowing for a potentially infinite number of parameters, i.e., an open-ended number of degrees of freedom in the model [203]. Thus, BNP models rely on random measures to induce an infinite-dimensional distribution over the parameter space, which we will define more rigorously in the next section.

2.3 Random Measures

A random measure M can be understood as a stochastic process indexed by a sigma algebra. Let $(\Phi, \mathcal{F}_\Phi)$ be a measurable space, where Φ is a set, and $\mathcal{F}_\Phi$ is a σ-algebra over Φ, for instance Φ is the real line and $\mathcal{F}_\Phi$ refer to the Borel sets. A random measure defines a distribution over measures on that measurable space; it corresponds to a collection of random variables $M(A) \in [0, \infty)$, one for each set $A \in \mathcal{F}_\Phi$. The expectation of a random measure is called the mean measure, which we denote by $\nu(A) \doteq \mathbb{E}[M(A)]$.

Completely random measures. A completely random measure (CRM) refers to a random measure such that the masses $G(A_1)$, $G(A_2)$, ... assigned to disjoint subsets A_1, A_2, ... $\in \mathcal{F}_\Phi$ by a draw G from the CRM are independent random variables [105]. The class of CRM includes important stochastic processes such as the Beta process (BP), Gamma Process, Poisson Process, and the stable

subordinator, which are often used as priors on a wide range of applications. A recent review of CRMs and their applications can be found in [117]. A CRM on some space $(\Phi, \mathcal{F}_\Phi)$ is characterized by a positive Lévy measure[8] $\nu(d\phi, d\pi)$ on the product space $\Phi \times \mathbb{R}^+$, with associated product σ-algebra $\mathcal{F}_\Phi \otimes \mathcal{F}_{\mathbb{R}^+}$. CRMs have a useful representation in terms of Poisson processes on this product space. Let $\Pi = \{(\phi_k, \pi_k) \in \Phi \times \mathbb{R}^+\}_{k=1}^\infty$ be a Poisson process on $\Phi \times \mathbb{R}^+$ whose rate measure is given by the Lévy measure $\nu(d\phi, d\pi)$. We denote this as $\Pi \sim \mathrm{PP}(\nu)$. A CRM can then be represented as an infinite sum of weighted atoms or sticks,

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \sim \mathrm{CRM}(\nu), \tag{2.12}$$

where $(\phi_k)_{k=1}^\infty$ is the set of atom locations, and $(\pi_k)_{k=1}^\infty$ is the set of atom weights, which do not necessarily sum to one. If $\sum_{k=1}^\infty \pi_k = 1$ holds, this CRM is then referred to as normalized random measure. Further details on CRMs can be found in [105].

Normalized random measures. Given a completely random measure $G \sim \mathrm{CRM}(\nu)$, we can construct a normalized random measure P based on an underlying probability measure on Φ as follows

$$P = \sum_{k=1}^{\infty} \frac{\pi_k}{\sum_{j=1}^\infty \pi_j} \delta_{\phi_k} \sim \mathrm{NRM}(\nu). \tag{2.13}$$

The most common example of normalized random measure is the DP, which can be obtained by normalization of the Gamma Process. Distributions over probability measures are of great importance in Bayesian statistics and machine learning, and normalized random measures have been used in many applications including topic modeling, image segmentation, or monitoring of genetic populations [200, 204].

Exchangeability and De Finetti's theorem Random measures are at the heart of BNP models via the *De Finetti's theorem*. Indeed, the theorem establishes a link between random measures and the so-called *exchangeability* property of the data. An infinitely exchangeable sequence of observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ is a sequence whose probability is invariant under finite permutations ρ of the first N elements, for all $N \in \mathbb{N}$ [54], i.e.,

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = p(\mathbf{x}_{\rho(1)}, \mathbf{x}_{\rho(2)}, \ldots, \mathbf{x}_{\rho(n)}),\ \forall N \in \mathbb{N}. \tag{2.14}$$

[8] In probability theory, a Lévy process is a stochastic process with independent, stationary increments, which can be viewed as the continuous-time analog of a random walk [105].

According to De Finetti's theorem, any infinitely exchangeable sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ can be written as a mixture of i.i.d. samples as follows:

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \int_{\phi} \prod_{n=1}^{N} \mathcal{Q}(\mathbf{x}_n|\phi) P(d\phi), \ \forall N \in \mathbb{N}, \tag{2.15}$$

where $\mathcal{Q}(\cdot|\phi)$ is a family of conditional distributions and P is a random measure over Φ called the *De Finetti mixing measure* (or directing measure).

For instance, if P is a discrete random measure drawn from a DP, and $\mathcal{Q}(\cdot|\phi) \doteq \mathcal{N}(\phi, \sigma_x \mathbf{I})$, we recover the example of the Gaussian mixture model described in Section 2.2. Also, if P is a DP, and $\mathcal{Q}(\cdot|\phi)$ is the discrete probability distribution described by the probability measure ϕ, we obtain a distribution over exchangeable partitions known as the Chinese restaurant process (CRP) [6]. As another example, if P is a BP and $\mathcal{Q}(\cdot|\phi)$ is a Bernoulli process, we obtain a distribution over exchangeable binary vectors; such construction defines the so-called Indian buffet process (IBP) [206]. In the following, we describe the two basic random measures that will be used along this thesis, together with their corresponding collapsed processes: the DP and the BP, giving rise to the CRP and IBP respectively.

2.3.1 Dirichlet Process

The Dirichlet process (DP) is a stochastic process whose realizations are random infinite discrete probability distributions [53]. Let G_0 be a probability random measure on the measurable space $(\Phi, \mathcal{F}_\Phi)$. A random probability measure G over $(\Phi, \mathcal{F}_\Phi)$ is said to be a DP if, for any finite measurable partition $(A_1, A_2, \ldots, A_r)$ of Φ, the random vector $\Big(G(A_1), G(A_2), \ldots, G(A_r)\Big)$ is distributed as a finite-dimensional Dirichlet distribution of the form

$$\begin{aligned} &\Big(G(A_1), G(A_2), \ldots, G(A_r)\Big) \\ &\sim \text{Dirichlet}\Big(\alpha G_0(A_1), \alpha G_0(A_2), \ldots, \alpha G_0(A_r)\Big). \end{aligned} \tag{2.16}$$

We then write $G \sim \text{DP}(\alpha, G_0)$, where G_0 is referred to as the base measure (which is the expected value of the process) and $\alpha \in \mathbb{R}^+$ is the concentration parameter, which plays the role of an inverse variance. Thus, the weak distribution of a DP, i.e., the set of all its finite-dimensional marginals follow a Dirichlet distribution. The first two cumulants of the DP are given by

$$\mathbb{E}[G(A)] = G_0(A), \qquad \text{and} \qquad \text{Var}[G(A)] = \frac{G_0(A)(1 - G_0(A))}{\alpha + 1}. \tag{2.17}$$

An explicit representation of a draw $G \sim \mathrm{DP}(\alpha, G_0)$ from a DP can be written as

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \tag{2.18}$$

where π_k are the atom weights, and $\phi_k \in \Phi$ are atom locations defined in the parameter space [188]. The representation in (2.18) shows that draws from a DP are atomic (discrete) with probability one.

As illustrated in Section 2.2, Eq. 2.18 defines an infinite mixture model, i.e., a mixture model with a countably infinite number of clusters. However, since the weights π_k decrease exponentially quickly, only a small number of clusters will be used to describe the data *a priori*. In fact, the expected number of mixtures grows logarithmically with the number of observations. In the DP mixture model, the actual number of clusters describing the data is not fixed, and can be automatically inferred from the data using the usual Bayesian posterior inference framework. The DP mixture model has been widely studied in the literature [12, 50, 124].

In the following, we describe other closely-related processes and explicit representations of the DP, including the stick-breaking process and culinary metaphor of the CRP. Explicit representations of stochastic processes directly describe a random draw from the stochastic process, rather than describing its distribution.

Stick-Breaking Representation

The discrete random measure G in (2.18) is uniquely determined by two infinite sequences, $\{\pi_k\}_{k\in\mathbb{N}}$ and $\{\phi_k\}_{k\in\mathbb{N}}$. As stated in [188], the stick-breaking representation of the DP generates these two sequences by drawing $\phi_k \sim G_0$ independently, and by drawing a set of auxiliary variables $(v_k)_{k=1}^{\infty}$ such that

$$v_k \sim \mathrm{Beta}(\alpha, 1), \qquad \text{and} \qquad \pi_k = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell). \tag{2.19}$$

Note that the sequence of atom weights $\{\pi_k\}_{k\in\mathbb{N}}$ constructed by (2.19) satisfies $\sum_{k=1}^{\infty} \pi_k = 1$ with probability one. We write $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$ if $\boldsymbol{\pi}$ is a random probability measure over the positive integers defined by 2.19 (GEM stands for the authors Griffiths, Engen and McCloskey) [153].

Figure 2.4 illustrates a sequential construction of the sequence $\{\pi_k\}_{k\in\mathbb{N}}$. Starting with a stick of unit length, at each iteration $k = 1, 2, \ldots, \infty$, a piece of relative length v_k is "broken off" (relative to the current length of the stick).

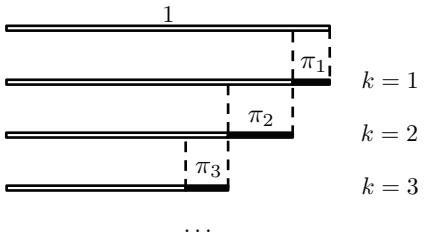


Figure 2.4: **Illustration of the stick-breaking construction for the DP.** Drawing $G \sim \mathrm{DP}(\alpha, G_0)$ is equivalent to drawing $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$ and $\phi_k \sim G_0$, $\forall k = 1, \ldots, \infty$ independently.

Chinese Restaurant Process

The CRP is typically found in the literature of DPs. This process defines a distribution on infinite partitions of the data [6], and takes the name from a standard culinary metaphor that vividly illustrates how the DP operates [204]. In this metaphor, the atom locations ϕ_k are referred to as "dishes" in a restaurant, and observations that are clustered together are viewed as customers sitting on the same table, therefore eating from the same dish. In this generative process, customers enter the restaurant one at a time, and they can either sit on an existing table, with probability proportional to the number of previous customers that are already sitting on that table, or open a new table, with probability proportional to α. In the latter case, they also sample a new dish from the prior, i.e., $\phi_{k^{\text{new}}} \sim H$.

The implicit representation of the DP, which is closely-related to the CRP, is the *Pólya urn scheme* [19]. Let $\theta_1, \ldots, \theta_N$ be a sequence of i.i.d. random variables distributed according to G. That is, all variables θ_n are conditionally independent given G, and hence exchangeable. The successive conditional distribution of θ_n given $\theta_1, \ldots, \theta_{n-1}$ takes the form

$$\theta_n | \theta_1, \ldots, \theta_{n-1}, \alpha, G_0 \sim \sum_{\ell=1}^{i-1} \frac{1}{i - 1 + \alpha} \delta_{\theta_\ell} + \frac{\alpha}{i - 1 + \alpha} G_0. \tag{2.20}$$

This expressions shows that θ_n has non-zero probability of being equal to one of the previous draws. This leads to a "rich gets richer" effect, in which the more often a point is drawn, the more likely it is to be drawn in the future. Let $\{\phi_k^\star\}_{k=1}^K$ denote a sequence containing the unique values of variables θ_n. By defining m_k as the number of values θ_n which are equal to $\phi_k^\star$, we can rewrite (2.20) as

$$\theta_n | \theta_1, \ldots, \theta_{n-1}, \alpha, G_0 \sim \sum_{k=1}^{K} \frac{m_k}{i - 1 + \alpha} \delta_{\phi_k^\star} + \frac{\alpha}{i - 1 + \alpha} G_0. \tag{2.21}$$

Note that, despite this "richer gets richer" effect, the probability of θ_n being drawn from G_0 (and hence being different to all previous values) is always positive, and proportional to α.

Eqs. 2.20 and 2.21 can be interpreted as a Pólya urn model, in which a ball θ_n is associated with a color $\phi_k^\star$. The balls are drawn from the urn equiprobably. When a ball is drawn, it is placed back in the urn together with a new ball of the same color. In addition, with probability proportional to α, a new atom (color) is created by drawing from G_0, and a ball of that new color is added to the urn. Alternatively, this process can also be illustrated using the same culinary metaphor as for the CRP. According to this metaphor, we consider a Chinese restaurant with an infinite number of tables. Each θ_n corresponds to a customer who enters the restaurant, while the distinct values $\phi_k^\star$ correspond to the tables at which the customers sit. The n-th customer sits at the table $\phi_k^\star$ with probability proportional to the number of customers m_k already seated there (in which case we set $\theta_n = \phi_k^\star$), or sits at a new table with probability proportional to α (therefore increasing K by one, drawing then $\phi_K^\star \sim G_0$, and setting $\theta_n = \phi_K^\star$).

Finally, let us explicitly relate the Pólya urn scheme to the CRP. For that, let us now introduce auxiliary random variables $z_1, z_2, \ldots, \infty$ to indicate the assigned atom index of each random draw, e.g., $\theta_n = \phi_{z_n} = \phi_k^\star$. Compared to the Pólya urn scheme in which we considered the conditional distribution of the *actual color* of each new drawn ball, we now consider the color *index* indicating that color. In other words, the CRP describes the sequential probability of each cluster assignment variable,

$$z_n | z_1, \ldots, z_{n-1}, \alpha, G_0 \sim \sum_{k=1}^{K} \frac{m_k}{i-1+\alpha} \delta_k + \frac{\alpha}{i-1+\alpha} \delta_{k_{new}}. \tag{2.22}$$

where $k = 1, \ldots, K$ correspond to the atom indexes of the already observed clusters (colors that have already appeared in the Pólya urn model or non-empty tables in the CRP), and k_{new} corresponds to a completely new index. The iterative process in Eq. 2.22 defines a distribution over infinite partitions. See Figure 2.5 for a sketch of the CRP.

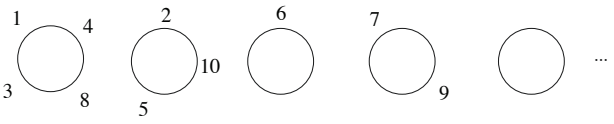


Figure 2.5: **Illustration of the Chinese restaurant process.** Circles correspond to tables in the restaurant, while numbers correspond to customers sitting on tables.

Hierarchical Dirichlet Process

The hierarchical Dirichlet process (HDP) is a well-known BNP prior which is useful for modeling grouped data [204]. The HDP is a distribution over a set of random probability measures. The process defines a set of random probability measures G_j (one for each group of data), and a global

random probability measure G. The global measure G is distributed as a DP with concentration parameter γ and base probability measure G_0, i.e., $G \sim \mathrm{DP}(\gamma, G_0)$. Hence, it can be written as

$$G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k}. \tag{2.23}$$

The random measures G_j are conditionally independent given G, and they are distributed as DPs with concentration parameter α and base probability measure G, i.e., $G_j \sim \mathrm{DP}(\alpha, G)$. Since G is a discrete probability measure with support at the points $\{\phi_k\}_{k\in\mathbb{N}}$, each probability measure G_j has support on the same set of points and, therefore, we can write

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}. \tag{2.24}$$

In other words, the atom weights[9] π_{jk} are different for each group j, but the atom locations are shared across groups j.

Let $(A_1, A_2, \ldots, A_r)$ be a measurable partition of Φ. Since $G_j \sim \mathrm{DP}(\alpha, G)$ for each j, we have by definition that the random vector $\Big(G_j(A_1), G_j(A_2), \ldots, G_j(A_r)\Big)$ is distributed as

$$\begin{aligned} &\Big(G_j(A_1), G_j(A_2), \ldots, G_j(A_r)\Big) \\ &\sim \mathrm{Dirichlet}\Big(\alpha G(A_1), \alpha G(A_2), \ldots, \alpha G(A_r)\Big). \end{aligned} \tag{2.25}$$

This will be useful to establish the connection between the weights π_{jk} and the weights of the global measure, ρ_k. Note that Eq. 2.24 defines an infinite mixture model for each group of observations j. Furthermore, all groups share the atom locations ϕ_k (which do not depend on j), and they also share statistical strength on the weights π_{jk}.

2.3.2 Beta Process

A BP is a CRM which is often used as a Bayesian nonparametric prior for sparse collections of binary features [206]. BPs are defined on an abstract measurable space $(\Phi, \mathcal{F}_\Phi)$ and have the property that the mass of any particular atom lies in the interval $[0, 1]$. The BP can be constructed based on a random measure $\Pi \sim \mathrm{PP}(\nu)$ of a Poisson Process defined on $(\Phi \times [0,1], \mathcal{F}_\Phi \times \mathcal{F}_{[0,1]})$, where ν refers to its mean measure $\nu(d\phi, d\pi) = \alpha\pi^{-1}(1-\pi)^{(\alpha-1)} d\pi P(d\phi)$ [145]. The first dimension of

[9] We use the notation $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots]$ to refer to the atom weights of the global probability measure G, and $\boldsymbol{\pi}_j = [\pi_{j1}, \pi_{j2}, \ldots]$ to refer to the atom weights of each group-level probability measure G_j.

the Poisson Process corresponds to atom location, and the second dimension to atom weight.

A draw G from a BP is a countably infinite collection of weighted atoms in a space Φ with probability one, whose weights lie in the interval $[0, 1]$ [117]. The distribution on these weights is governed by two parameters (similarly to the DP): a concentration parameter $\alpha > 0$ and a base measure H, and we write $G \sim \mathrm{BP}(\alpha, H)$. But in contrast to the DP which provides a probability measure, the total measure $G(\Phi) \neq 1$ with probability one. Beta processes are often used in a hierarchical prior as parameters for a Bernoulli process (BeP), which we will denote as $X|G \sim \mathrm{BeP}(G)$. When the BP is marginalized out, we obtain the IBP [206].

Stick-Breaking Construction

The stick-breaking construction of the BP in [202] is an useful representation for some inference algorithms. In this construction, a sequence of independent Beta-distributed random variables $\{v_k\}_{k\in\mathbb{N}}$ are used to obtain the atom weights $\{\pi_k\}_{k\in\mathbb{N}}$. These are generated according to

$$v_k \sim \text{Beta}(\alpha, 1), \qquad \pi_k = \prod_{\ell=1}^{k} v_\ell, \tag{2.26}$$

resulting in a decreasing sequence of probabilities π_k. This construction can be understood with the stick-breaking process illustrated in Figure 2.6. Starting with a stick of length 1, at each iteration $k = 1, 2, \ldots$, a piece is broken off at a point v_k relative to the current length of the stick. The variable π_k corresponds to the length of the stick just "broken off", and the other piece of the stick is discarded.

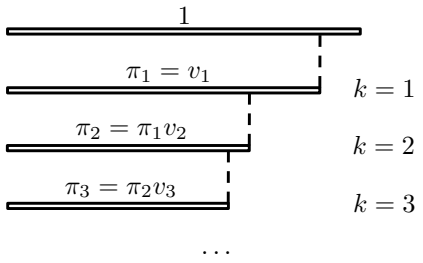


Figure 2.6: **Illustration of the stick-breaking construction for the BP.**

Indian Buffet Process

The IBP is a stochastic process defining a probability distribution over equivalence classes of sparse binary matrices with a finite number of rows and an unbounded number of columns [72]. Although the number of columns is potentially infinite, only a finite number of those will contain non-zero entries due to the finite nature of the observed data. Another important property of IBP-generated

matrices is that they are exchangeable both in rows and columns. The IBP can be derived taking the limit as $K \to \infty$ of a finite binary matrix $\mathbf{Z} \in \{0,1\}^{N \times K}$, whose elements z_{nk} are distributed according to:

$$\pi_k \sim \text{Beta}(\alpha/K, 1), \qquad z_{nk} \sim \text{Bernoulli}(\pi_k), \tag{2.27}$$

where π_k is the probability of observing a non-zero value in column k, $z_{n\bullet}$ is the n-th row for sample n and $z_{\bullet k}$ is the k-th column of matrix $\mathbf{Z}$. We say that a feature k is active for sample n if $z_{nk} = 1$. When $K \to \infty$, the above process is equivalent to the IBP. We denote:

$$\mathbf{Z} \sim \text{IBP}(\alpha), \tag{2.28}$$

where α is the concentration parameter controling the a priori activation probability of new features. The expected number of active features per row is distributed according to $\text{Poisson}(\alpha)$ and the actual number K of features with non-zero elements is distributed as $\text{Poisson}\big(\alpha \cdot \sum_{i=1}^{N} \left(\frac{1}{i}\right)\big)$. The parameter α has thus an effect on both, the a priori number of columns in $\mathbf{Z}$ and degree of sparseness of the matrix, e.g., a bigger value for α results in a higher number of expected latent features and more active features per row a priori.

As discussed previously, the underlying De Finetti's representation of the IBP is a mixture of BePs directed by a BP as follows:[10]

$$G \sim \text{BP}(1, \alpha, H) \tag{2.29}$$

$$\mathbf{Z}_{n\cdot} \sim \text{BeP}(G), \tag{2.30}$$

where G is the directing measure, and H is the probability base measure for the BP [206]. The IBP can also be directly obtained by combining the previously described stick-breaking construction with an infinite collection of Bernoulli-distributed random variables (one for each atom) as:

$$\begin{aligned} v_k &\sim \text{Beta}(\alpha, 1), \quad \pi_k = \prod_{i=1}^{k} v_i, \\ z_{nk} &\sim \text{Bernoulli}(\pi_k). \end{aligned} \tag{2.31}$$

Similarly to the culinary metaphor for the CRP, there exists a sequential stochastic process that defines a conditional distribution $p(z_{nk}|\mathbf{Z}_{-nk})$ which can be illustrated using a new culinary metaphor that gives the name to this stochastic process. Imagine an Indian restaurant whose buffet

[10] Eq. 2.29 and 2.30 employ a common slight misuse of notation by ignoring the features' position of the beta and Bernoulli processes.

consists of infinitely many dishes arranged in a line. N customers enter the restaurant sequentially. The first customer starts at the left of the buffet and takes a serving from each dish, stopping after a $\text{Poisson}(\alpha)$ number of dishes, as his plate becomes overburdened. The n-th customer moves along the buffet and samples dishes in proportion to their popularity, serving himself with probability $\frac{m_k}{n}$, where m_k is the number of previous customers who have sampled dish k. Having reached the end of all previously sampled dishes, the n-th customer then tries a $\text{Poisson}(\frac{\alpha}{n})$ number of new dishes. This construction give us $\mathbf{Z}$ matrices whose features are ordered decreasingly according to their feature activation probability, although any feature permutation will have the same probability given the exchangeability property over columns. The actual distribution over the equivalence matrix $\mathbf{Z}$ is independent of the arrival order of customers entering the restaurant [202].

2.4 Inference Methods

The main computational problem in BNP modeling (as in most of Bayesian statistics) is computing the posterior distribution. For most interesting models, the posterior is computationally not tractable (not available in closed form), thus requiring inference algorithms to compute an approximation. The two most popular inference approaches are MCMC and variational inference techniques.

MCMC methods refer to iterative algorithms in which groups of latent variables are sampled given all the rest (such distributions are called the conditional distributions). This approach has asymptotic theoretical guarantees: it consists in defining a Markov chain on the different states of the hidden variables whose asymptotic distribution (equilibrium distribution) is known to converge to the posterior distribution. Samples obtained from this Markov chain will eventually correspond to samples from the posterior as the number of iterations tends to infinity [177]. A simple form of MCMC sampling is Gibbs sampling, where the Markov chain is constructed by considering the conditional distribution of each hidden variable given the rest of hidden variables and the observations. The CRP construction is particularly amenable to Gibbs sampling inference, as obtaining these conditional distributions is straightforward. A detailed survey of Gibbs sampling for inference in DP mixture models can be found in [140]. Gibbs sampling for the IBP is described in [72].

In contrast, variational inference resorts to transforming the inference problem into an optimization task [102]. The idea here is to approximate the posterior distribution by a simpler family of distributions and searching for the member of that family that is closest to it (according to the Kullback-Leibler divergence). This is also equivalent to maximizing a certain lower bound on the marginal distribution of the data, called the "Evidence Lower Bound" (ELBO). Unlike MCMC methods, vari-

ational inference algorithms are not guaranteed to recover the posterior, but they are typically faster than MCMC, and convergence assessment is straightforward. These methods have been applied to DP mixture models [21] and IBP latent feature models [42].

2.5 Summary

In this chapter, we have given a gentle introduction to Bayesian nonparametric and intuition, including concepts such as exchangeability, De Finetti's theorem, completely random measures and normalized random measures. We have presented the basic building blocks (stochastic processes) which will be used within this thesis, i.e., the Dirichlet process, and the Beta process, as well as the additional processes resulting from integrating the De Finetti's mixing random measure in a Dirichlet process and a hierarchical Beta process–Bernoulli process construction: the Chinese restaurant process and the Indian buffet process. Finally, we have given an overview of the most common inference approaches to fit BNP models, i.e, to approximate the posterior distribution. The remaining chapters are devoted to our contributions.

3

Atom-Dependent Dirichlet Process for Marathon Modeling

This chapter presents a novel application of Bayesian nonparametrics (BNPs) for density estimation of stratified data, with application to data from marathon runners. In particular, we make use of the dependent Dirichlet process (DDP) [125], which is a powerful tool that encompasses the Dirichlet process (DP) and the hierarchical Dirichlet process (HDP). However, the DDP is very general and it cannot be directly applied to data without additional constraints. Here, we specify a way to tie the parameters across groups using a Gaussian process (GP) [172], thus making the DDP a practical prior for our problem at hand. Additionally, we rely on the HDP to model intermediate running times for each runner, uncovering different running patterns within athletes. This model is also used to predict the finishing time in the race. Finally, we relate the proposed model to the literature of infinite mixture of experts (IMoE) in the context of non-linear regression [171].

3.1 Introduction

Fairness is becoming an important requirement for certain machine learning (ML) systems to have, some examples include non-gender-biased recommendation systems or unbiased decision-systems regarding racial discrimination. Our objective in this chapter is to compare in a fair manner the finishing time of runners having different age and gender. Currently, most popular marathons award entry to participants by their best marathon in the previous 12 months. For example, in Boston it is the only way that a participant can gain entry to the race, while other paths are available in New York, Chicago or London.[1] Our objective is to propose a methodology that can be used to equalize entry requirements for different marathons, which vary considerably for one event to the next, as there is no widely accepted standard method to specify them. Furthermore, the world master of athletics (WMA) has an age-grading system for equalizing the finishing time according to the age and gender of athletes.[2] They lobby for this measure to be taken into consideration for selecting the winners of each race, even though that procedure is based on world records, i.e., outliers that may not be very representative, or even realistic, for most races. Our method also provides an alternative way to reward runners fairly regardless of their age and gender.

Our approach consists in adapting a single-p dependent Dirichlet process (sp-DDP) to cluster the finishing time for each runner according to his/her age and sex [125]. We propose a Gaussian process to control how the clusters (representing marathon finishing time) change from one group to the next (different ages or gender). We find that the means of these clusters are directly comparable to the marathon entry requirements and the age-graded tables from the WMA. Additionally, direct comparisons for any finishing time are straightforward, since we find a full distribution for all ages and both genders. The sp-DDP can simultaneously deal with different races and/or the same race on different years, providing a unified ranking for all the races that may differ on elevation profile, temperature or humidity.

3.2 Dependent Dirichlet Processes

The DDP is a generalization of the DP that can be applied for clustering of groups of data [124]. For each group j, we have an infinite mixture model of the form

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_{jk}}, \tag{3.1}$$

[1] Complete information for each marathon event can be found in their respective web pages.
[2] World Master of Athletics webpage: `http://www.mastersathletics.net/`.

where G_j is a group-specific random measure, and the atom weights π_{jk} and atom locations ϕ_{jk} follow stochastic processes on the covariate space j. In the following, we will describe two particularizations of the DDP that we use throughout this chapter: the HDP and the sp-DDP.

Hierarchical Dirichlet process. The HDP, which was already introduced in Section 2.3.1, is a particular DDP to cluster groups of data that share the exact same mixture components [204]. It is equivalent to a set of DPs, one for each group of data, in which all DPs share the same base distribution which is itself drawn from a DP. Mathematically, we first draw a base distribution from a DP as $G_0 \sim \text{DP}(\gamma, H)$, where $G_0 = \sum_{k=1}^{\infty} v_k \delta_{\phi_k}$, and for each group j we draw a distribution from a DP using G_0 as the base distribution, i.e., $G_j \sim \text{DP}(\alpha, G_0)$. This construction ensures that all the random measures G_j share the same atom locations given by G_0, since G_0 is itself a discrete probability distribution. Since each atom corresponds to a cluster, cluster parameters ϕ_k are shared across all groups. Each G_j admits a representation as

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}, \tag{3.2}$$

where the atom locations ϕ_k do not depend on the group j. Furthermore, this method allows groups to share statistical strength via the atom weights v_k of the base distribution G_0. Indeed, the vector of weights for each group j can be obtained as $\boldsymbol{\pi}_j \sim \text{DP}(\alpha, \boldsymbol{v})$.

In the corresponding culinary metaphor, the HDP can be explained with a Chinese restaurant franchise (CRF), in which there is a collection of restaurants, and dishes are shared across restaurants. However, the popularity of each dish, i.e., the corresponding atom weight, is different in each of the restaurants [204].

Single-p dependent Dirichlet process. The sp-DDP is another DDP, which works in a complementary fashion to the HDP. In this case, atom weights are shared across groups while atom locations are allowed to vary. In terms of the often used culinary metaphor for DPs [61], while the HDP shares the dishes across restaurants but allows a different dish popularity in each of the restaurants, the sp-DDP shares the dish popularity across restaurants but allows the dishes to vary slightly, in order to better fit each group of customers (see Figure 3.1 for a comparative sketch). The latter would be a peculiar CRF in which the popularity of tables is matched one-to-one across restaurants, with the served dish in each linked table slightly customized in each restaurant (e.g., different ingredients, cooking time or local taste).

In the sp-DDP, the latent measure for each group j can be expressed as

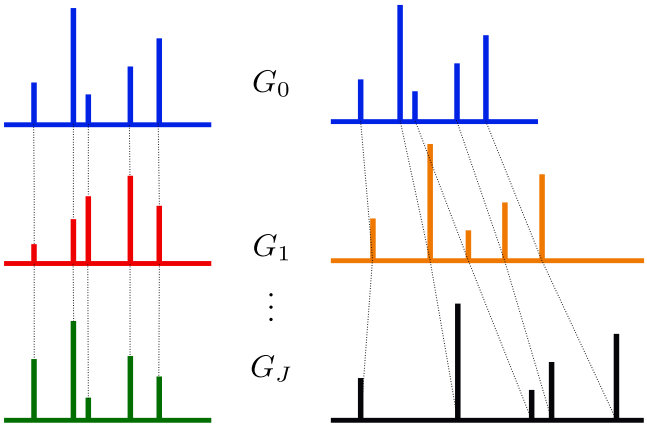


Figure 3.1: **Comparison of two Dependent Dirichlet processes.** The HDP is at the left, sp-DDP is at the right. The first one shares atom locations, while the second one shares mixture weights.

$$G_j = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_{jk}}, \tag{3.3}$$

where the vector $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$ contains the mixture weights and ϕ_{jk} are the atom locations. The sp-DDP does not specify how to tie the atom locations ϕ_{jk} across groups for each k. This step is critical, as it conditions the performance of the model. We put forward in Section 3.3 another stochastic process for this purpose.

3.3 Our Approach

3.3.1 Atom-Dependent Dirichlet Process Mixture Model

Our model is based on the sp-DDP prior, as it allows comparing the shape of different distributions while keeping the corresponding quantiles fixed. In the context of the marathon, we use it to obtain a fair comparison between groups of runners regardless of their age or gender. Runners are grouped together according to their age and gender, yielding J different groups. In our infinite mixture model, we cluster the runners of all groups with a potentially unbounded number of clusters. Each cluster k presents a fixed percentage of runners given by π_k, with a stochastic process linking the atom locations, i.e., the mean finishing time. This construction has the potential to provide a direct comparison for the finishing time in each group j. However, as the sp-DDP is a very general prior, we need to define the likelihood and the stochastic process in a way that is insightful about the marathoner's finishing time.

We denote each marathoner finishing time as x_{ji}, where $j = 1, \ldots, J$ indexes the group and i runs over marathoners, and we assume a Gaussian likelihood for the finishing time x_{ji}:

$$x_{ji}|c_{ji} = k, \mu_k, \theta_j, \sigma_x^2 \quad \sim \quad \mathcal{N}\left(x_{ji}|\mu_k + \theta_j, \sigma_x^2\right). \tag{3.4}$$

Here, μ_k denotes the global mean for cluster k, θ_j is the shift associated to group j, c_{ji} represents the cluster assignment associated to observation x_{ji}, and σ_x^2 is the variance of the Gaussian distributions. Hence, we use a cluster-specific parameter μ_k to describe cluster k, but we allow deviations from this value due to age or gender: this effect is modeled by θ_j.

The key aspect that makes the sp-DDP useful for comparing different age-gender groups is the stochastic process that governs θ_j. We would like this value to vary smoothly across ages, and therefore we choose a zero-mean Gaussian process prior for it:

$$\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right), \tag{3.5}$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_J]^\top$, and for the covariance function we use the standard choice, i.e., the squared exponential kernel given by

$$(\boldsymbol{\Sigma}_\theta)_{\ell j} = \sigma_\theta^2 \cdot \exp\left(-\frac{(\ell - j)^2}{2\nu^2}\right) + \kappa\delta(\ell - j), \tag{3.6}$$

where ℓ and j represent two different age groups, σ_θ^2 accounts for the variance, ν controls the degree of correlation between age groups ℓ and j, and κ is a jitter factor to avoid numerical instabilities. We use an independent Gaussian process for each gender. There are other alternatives, see [172] for a comprehensive introduction for valid covariance functions. In our case, the squared exponential kernel is a smooth kernel (infinitely differentiable) that captures the correlation between the different age groups. We place a Gaussian prior over the cluster means μ_k and an inverse gamma prior over the variance σ_x^2, i.e.,

$$\mu_k \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right), \qquad\qquad \sigma_x^2 \sim \mathcal{IG}\left(a, b\right), \tag{3.7}$$

where μ_0, σ_0^2, a and b are hyperparameters of the model. The value of σ_0^2 is assumed to be much larger than σ_θ^2, so that the first one controls the overall finishing time for the clusters (hours), whilst σ_θ^2 controls the differences between groups due to different ages (minutes).

Finally, we place the following priors over the assignment variables c_{ji} and cluster weights $\boldsymbol{\pi}$:

$$c_{ji}|\boldsymbol{\pi} \sim \boldsymbol{\pi}, \qquad\qquad \boldsymbol{\pi}|\alpha \sim \text{GEM}(\alpha), \tag{3.8}$$

which completes the specification of the generative model and GEM stands for the stick-breaking prior by Griffiths, Engen and McCloskey, as defined in [152]. We refer to this sp-DDP prior, together

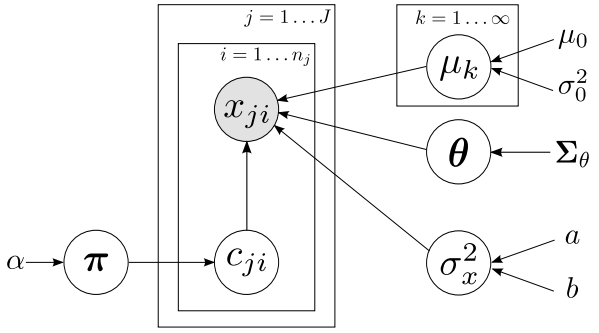


Figure 3.2: **Graphical representation of the basic ADDP mixture model.** Grey circles represent observed variables, white circles are hidden random variables. Plates refer to duplicated random variables.

with the likelihood model in Eq. 3.4 and the GP for θ_j, as the Atom-dependent Dirichlet process (ADDP) mixture model. A graphical representation for the ADDP is depicted in Figure 3.2. This model is similar to the ANOVA-DDP prior in [37]: in our case, we rely on a GP to control the dependency between the shift variables θ_j, while the ANOVA-DDP prior relies on independent priors. This allows for smooth variations of the cluster means with the age.

We assume a common variance σ_x^2 for all the clusters, because it provides an ordering of the clusters, which is necessary for comparing the finishing times. Allowing for different variances for each component should provide a more accurate (in the sense of better density estimation) description of the finishing times, but a less interpretable and less actionable representation, as runners assigned to Gaussian components with different variances are not directly comparable. We have not placed a joint prior for the cluster means and variances through the normal-inverse gamma distribution, since separate priors might have better properties for density estimation [76] and allow for faster Gibbs sampling inference.

3.3.2 Further Model Extensions

Age-gender interaction. The basic ADDP mixture model considers male and female runners independently, assuming independent shift delays θ_j between both genders j (i.e., we use a block-diagonal covariance matrix). A natural extension of the model consists in introducing an additional gender factor δ and some age-gender interaction factors ω_j in order to capture the correlation between male and female athletes. In such model, the shift delays θ_j are shared for both men and women, and j indexes different *age* groups instead of *age-gender* groups. We refer to this model as the age-gender interaction ADDP model. The generative model can be written as follows:

$$x_{ji}|c_{ji} = k, g_{ji}, \mu_k, \theta_j, \sigma_x^2, \delta, \omega_j \sim \mathcal{N}\left(x_{ji}|\mu_k + \theta_j + \mathbb{1}[g_{ji} = 1](\delta + \omega_j), \sigma_x^2\right), \tag{3.9}$$

$$c_{ji}|\boldsymbol{\pi} \sim \boldsymbol{\pi} \qquad \boldsymbol{\pi}|\alpha \sim \mathrm{GEM}(\alpha) \qquad \sigma_x^2 \sim \mathcal{IG}\left(a, b\right) \tag{3.10}$$

$$\begin{aligned} \mu_k &\sim \mathcal{N}\left(\mu_0, \sigma_0^2\right) & \boldsymbol{\theta} &\sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right) \\ \delta &\sim \mathcal{N}\left(0, \sigma_g^2\right) & \boldsymbol{\omega} &\sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\omega\right) \end{aligned} \tag{3.11}$$

where the indicator variables g_{ji} differentiate male ($g_{ji} = 0$) and female ($g_{ji} = 1$) runners, δ is the gender effect, and $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_J]^\top$ contains the age-gender interaction factors influencing the likelihood of female runners. Additionally, σ_g^2 and $\boldsymbol{\Sigma}_\omega$ are hyperparameters of the model.

Multiple races. We could apply the previous models to finishing times from different races or years, but we might obtain unexpected results since different races can present unalike conditions due to temperature, elevation profile, humidity, pull of runners, etc. In this section, we present an useful extension that deals with different races all together and allows drawing comparisons between these races. In order to deal with this, we extend the basic ADDP model using varying weights across races. This leads to a hierarchical atom-dependent Dirichlet process (H-ADDP) model, in which cluster weights are allowed to change across races, and cluster parameters are allowed to change across age-gender groups. We refer to this model as the hierarchical H-ADDP model. Figure 3.3 shows a graphical representation of this extended model, with the following likelihood and priors:

$$x_{rji}|c_{rji} = k, \mu_k, \theta_j, \sigma_x^2 \sim \mathcal{N}\left(x_{rji}|\mu_k + \theta_j, \sigma_x^2\right), \tag{3.12}$$

$$\begin{aligned} \mu_k &\sim \mathcal{N}\left(\mu_0, \sigma_0^2\right), & c_{rji}|\boldsymbol{\pi}_{r\bullet} &\sim \boldsymbol{\pi}_{r\bullet}, \\ \boldsymbol{\theta} &\sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right), & \boldsymbol{\pi}_{r\bullet}|\boldsymbol{v}, \alpha &\sim \mathrm{DP}(\alpha, \boldsymbol{v}), \\ \sigma_x^2 &\sim \mathcal{IG}\left(a, b\right), & \boldsymbol{v}|\gamma &\sim \mathrm{GEM}(\gamma), \end{aligned} \tag{3.13}$$

where r indexes the different races, γ is the upper level concentration parameter, $\boldsymbol{\pi}_{r\bullet} = (\pi_{rk})_{k=1}^\infty$ are the mixture weights for race r, $\boldsymbol{v} = (v_k)_{k=1}^\infty$ are the global weights for all races.

One simple way to interpret this model is by conditioning on a particular race or an age-gender group. If we only have data from a single race, we recover our original ADDP model. If we only have data from a single age-gender group, we recover an HDP, i.e., the cluster components are shared, but the mixing proportions are different.

Cluster-dependent shifts. We now present another extension of the model concerning the shifts $\boldsymbol{\theta}$. In the model described above, the delay θ_j only depends on the age and gender, which implies that the shift is the same for all clusters k, no matter whether they are fast or slow runners. However,

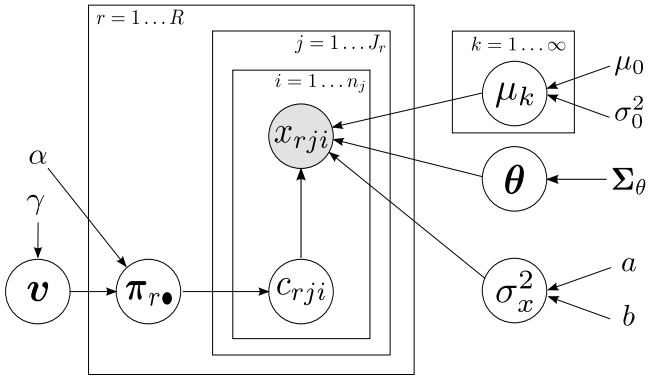


Figure 3.3: **Graphical representation of the H-ADDP mixture model.** Cluster weights change across races, whereas cluster means change across age-gender groups.

we can also consider cluster-dependent shifts. This allows us to capture different shift evolutions across age/gender depending on the speed of runners. That is, instead of having a single delay θ_j for each group, we consider a different delay θ_{jk} for each cluster and group. Each vector $\boldsymbol{\theta}_{\bullet k} = [\theta_{1k}, \dots, \theta_{Jk}]^\top$ follows its own Gaussian process with mean $\boldsymbol{\mu}_\theta$ and covariance matrix $\boldsymbol{\Sigma}_\theta$:

$$\boldsymbol{\theta}_{\bullet k} \sim \mathcal{N}\left(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta\right). \tag{3.14}$$

3.4 Results

In the following experiments, we apply the described models to the New York City marathon[3] for 6 different years, between 2006 and 2011. This database consists of $249,899$ runners in total. We additionally compare the NYC marathon data to the marathons of Boston[4] and London[5] for 2010 and 2011, including $117,255$ additional runners. In order to test the resulting models, we set aside a test set with 20% of the participants for each race and age/gender group, ensuring that the age and gender proportions are the same in both train and test sets.

Concerning the hyperparameters for the basic ADDP mixture model in Section 3.3, we have set: $\sigma_\theta^2 = 0.05$, $\nu = 10$, $\kappa = 10^{-6}$, $a = 1$, $b = 1$, $\mu_0 = 5$ and $\sigma_0^2 = 1$. The value of μ_0 and σ_0 are set so that 2-hour marathoners are within 3 standard deviations and 9-hour marathoners (typical cut-off time) are not unheard of. The ratio between σ_θ and σ_0 is approximately 1 to 4, so the former is of the order of 15 minutes, while the latter is of the order of one hour. Therefore, σ_0 controls the overall finishing time, while σ_θ controls the differences among age groups. The values of a and b control the variance of each Gaussian component and are set so that values of less than an hour, but not too small,

[3]NYC marathon data at `http://www.tcsnycmarathon.org/about-the-race/results`
[4]Boston marathon data at `http://www.stat.unc.edu/faculty/rs/Bostonwebpage/readme.html`
[5]London marathon data at `http://www.virginmoneylondonmarathon.com/`

are more likely. ν was defined in Eq. 3.6 and controls the correlation between the mean finishing time of runners having same gender and different ages, so that they do not deviate significantly. The value of κ measures the error in the recorded time and in this case it prevents numerical instabilities. In the age-gender interaction ADDP model, we choose the same hyperparameters as in the basic ADDP model, and the additional hyperparameters are set as $\sigma_g^2 = 0.05$ (which corresponds to the same prior variance as for θ_j) and $\mathbf{\Sigma}_\omega = \frac{1}{2}\mathbf{\Sigma}_\theta$.

For the HDP model in Section 3.5.1, we set $\epsilon = 0.2\ \text{km}^{-1}$ and $\tau = 5000$, because the differences between the relative time spent at each 5-km interval are typically small. Finally, we place a Gamma prior with shape 1 and scale 10 over the concentration parameters α and γ, and sample their values following [50]. These hyperpriors are chosen in order to avoid the creation of too many spurious clusters. Due to the huge amount of data, results are not very sensitive to hyperparameter values, as long as they are not set to completely misleading and unrealistic values. In our application at hand, we can actually set (and be able to explain) the values of the hyperparameters using all our prior knowledge, as detailed above, which is important in order to incorporate the known information and allow for the expected variances.

Posterior inference for all simulations is based on Gibbs sampling. Following Algorithm 8 in [140], we do not integrate out the hidden variables, and we propose 10 new clusters at each iteration. In our results in Section 3.4.1, we report the values of the hidden variables (means and shift delays) averaged for the last $10,000$ iterations after running the sampler for $50,000$ iterations. For the per-cluster variables, we carry out the averaging procedure to account for potential label switching.

3.4.1 Density Estimation

To model the finishing time of runners in the six considered NYC marathons, we compare our basic ADDP mixture model described in Section 3.3 to a standard HDP mixture model with Gaussian likelihood. For both models, we report 11 clusters, with $\sigma_x = 5.8$ minutes for the basic ADDP model, and $\sigma_x = 8.2$ minutes for the HDP model. Figure 3.4 compares the overall density estimate given by both models with the empirical histogram for a particular group of runners (we choose all forty year-old male finishers for the plot). Both the ADDP and HDP models perform similarly in terms of density estimation, and both provide comparable test log-likelihood values, in particular, -0.0618/sample for the ADDP and -0.1053/sample for the HDP.

Note also that the empirical histogram in Figure 3.4 presents one narrow peak that is not fully captured by the HDP nor the ADDP model. This peak, just under 4 hours, and the valley right afterwards are due to some runners trying to finish (and succeeding) a sub-4-hour marathon, roughly

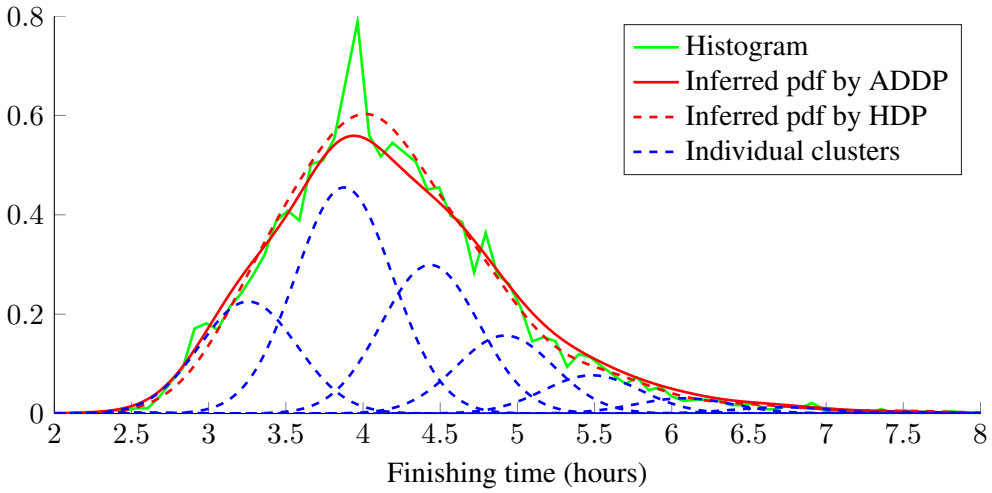


Figure 3.4: **Density estimation capacities for the basic ADDP and HDP model.** The histogram corresponds to the population of forty-year-old male runners, which is the largest age-gender group. The red curves are the probability density functions inferred by the basic ADDP and HDP models. The blue dotted lines represent the individual clusters inferred by the ADDP model.

9-minute-mile pace. This is a psychological effect that has limited interest for us, since it is not indicative of runners' inherent performance. Using cluster-specific values for the variance σ_x^2 would yield better density estimation, even capturing this peak, but it would fail to provide any comparison between distributions. Here, we are interested in ordering runners into clusters for comparison, which is achieved by having a shared value of σ_x^2 for all clusters.

3.4.2 Impact of Age, Gender and Race

We now use the age-gender interaction ADDP model described in Subsection 3.3.2. In addition to its density estimation capacity, the ADDP has an additional descriptive strength, since it can show the impact of age and gender on runners performance straightforwardly through inference of the age delays θ_j, gender factor δ and age-gender interaction factors ω_j. Figure 3.5 shows the average proportion of runners in each cluster, as well as the inferred cluster means μ_k, for both the HDP and the ADDP. Runners aged above 69 are not reported because there are too few of them.

The HDP results are not easy to interpret, except for the first three clusters that show the time degradation with age, because we do not know the cumulative percentage as the finishing time is increased. In contrast, the ADDP model is easier to interpret. The first cluster contains the "Olympic" quality runners for all ages, if Olympics were held for each age group (less than 1% of the runners). The second cluster has the competitive runners (about 13% of the runners), the third cluster has the

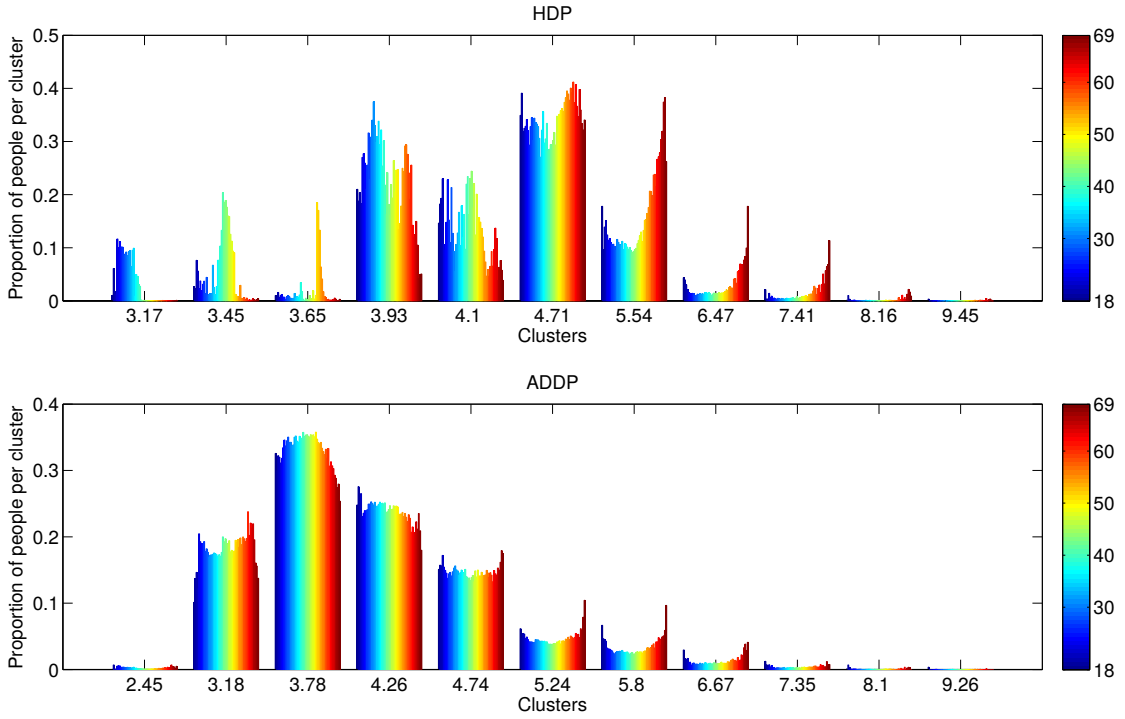


Figure 3.5: **Age proportions per cluster.** The HDP is at the top and the age-gender interaction ADDP is at the bottom. Male and female runners are shown together. The colors design the age of the runners, from 18-year-old in blue to 69-year-old in dark red. For the ADDP model, the cluster labels correspond to $\mu_k + \theta_j$, with the value of θ_j corresponding to the 28-year-old male runners, the shifts for other age or gender can be found in Table 3.1.

age	18	19								
men	8.00	6.34								
women	33.60	31.89								
age	20	21	22	23	24	25	26	27	28	29
men	4.88	3.62	2.53	1.66	0.98	0.49	0.18	0.04	0.00	0.10
women	30.43	29.21	28.19	27.42	26.87	26.53	26.38	26.41	26.53	26.79
age	30	31	32	33	34	35	36	37	38	39
men	0.28	0.51	0.79	1.09	1.39	1.71	2.03	2.34	2.65	2.98
women	27.13	27.50	27.87	28.27	28.65	29.03	29.39	29.74	30.08	30.46
age	40	41	42	43	44	45	46	47	48	49
men	3.30	3.66	4.11	4.61	5.25	5.99	6.85	7.90	9.09	10.47
women	30.82	31.23	31.78	32.41	33.21	34.13	35.24	36.58	38.07	39.81
age	50	51	52	53	54	55	56	57	58	59
men	12.01	13.74	15.66	17.74	20.00	22.38	24.91	27.51	30.21	32.95
women	41.73	43.88	46.23	48.75	51.47	54.29	57.26	60.27	63.36	66.44
age	60	61	62	63	64	65	66	67	68	69
men	35.70	38.44	41.16	43.80	46.35	48.77	51.06	53.17	55.09	56.81
women	69.50	72.51	75.46	78.26	80.93	83.42	85.73	87.83	89.69	91.31

Table 3.1: **Averaged values of θ_j for men (or $\theta_j + \delta + \omega_j$ for women) for all age groups in minutes.**

standard marathoners (about 33% of all runners), and so on. The x-axis in Figure 3.5 provides the value of $\mu_k + \theta_j$ for 28 year-old male runners and the degradation for other ages and sex is shown in Table 3.1. Using the plot and the table, we can know the proportion of runners in each group and how much extra time they need compare to the fastest group.

Figure 3.6 shows the value of the inferred cluster means μ_k plus and minus one standard deviation, shifted according to θ_j, δ, and ω_j for both men and women. We only depict the two fastest clusters, and compare the corresponding values of the finishing time with the entry requirements of different marathons and the WMA records. Best performance for females and males is predicted, respectively, at 26 and 28 years old, which is consistent with [186]. The plateau afterwards illustrates a stable period of performance between the 30's and 40's for both genders.

All plots behave in a similar way. This is what we meant in the introduction by that *the first insights should not be foreign to us*, so experts in marathon modeling can take other conclusions as plausible. Now, we focus on what is different. The most striking difference is how the entry levels penalize younger male runners, specially runners under 25. To be fairer to the youngest runners, their entry time should be raised (in about 7 minutes compared to the 30 years-old). The Boston marathon entry level is perfectly aligned with our second cluster for 40+ years old men and almost perfectly match the female second cluster, except for the runners under 23. There is a penalty of about 4-7 minutes for runners aged 25-39 and between 7-14 minutes for 18-24 year-old runners. The entry times slightly favor the 45-50 year-old runners.

The London marathon also penalizes excessively runners in their fifties compared to those in their forties and sixties, which seems odd.[6] It is also clear that over 50 (or even 45), the degradation of the finishing times per year is significant enough to merit a finer scale to guarantee entry times (this may also apply for 18-23 years-old). For example, a runner of 60 years old is doing almost 15 minutes less than a runner of 64 (which is a very long time in any marathon). Finally, the WMA curve for men penalizes the older male runners, while for older female runners it seems to have a similar trend than the first cluster of the ADDP model. For younger runners, the difference between the typical women in the olympic cluster and the WMA is larger than the difference between the typical men in the olympic cluster and the WMA.

Figure 3.7 shows the finishing time gap between women and men. The gap seems to be of about 30 minutes and slightly increasing with age. There are very few runners over 65 for the final decay to be statistically significant. We can come up with two different plausible explanations, but we do not have data to confirm whether this empirical effect is due to any of them or to some other unknown

[6] the entry time in London is only for UK residents and it is not a common way to get in the race, which might explain the weird pattern.

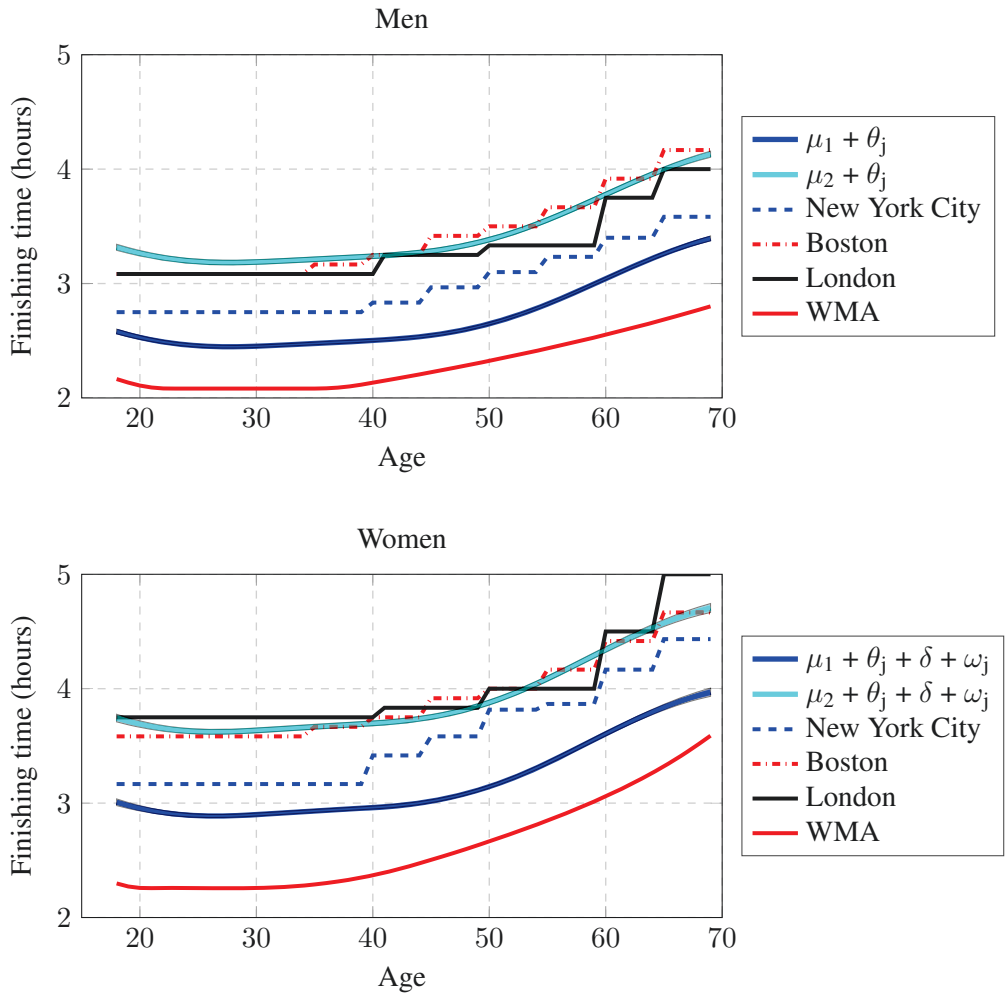


Figure 3.6: **Inferred cluster means and entry requirements.** Comparison of the inferred cluster means, i.e., $\mu_k + \theta_j$ with $k \in \{1, 2\}$, with the entry requirements (runners below the curve can qualify) for New York City, Boston and London marathons. (Top) Men. (Bottom) Women.

factor. The degradation with age between women and men might be due to physiological factors, i.e., women age differently than men for long distance running, or it might be due to socio-economical factors, i.e., women above 30 cannot train as much as men do.

Comparison across Multiple Races

We compare the NYC marathons to the ones in Boston and London, using the H-ADDP model described in Section 3.3.2. We consider both the 2010 and 2011 marathons, and we split the runners into *age groups* instead of using their actual age because we do not have this data available for London marathon. Figure 3.8 shows the inferred values of the per-race weights π_{rk}. The values for $\mu_k + \theta_j$ in the x-axis are those of the 45-49 male runners and the value of $\sigma_x = 19$ minutes.

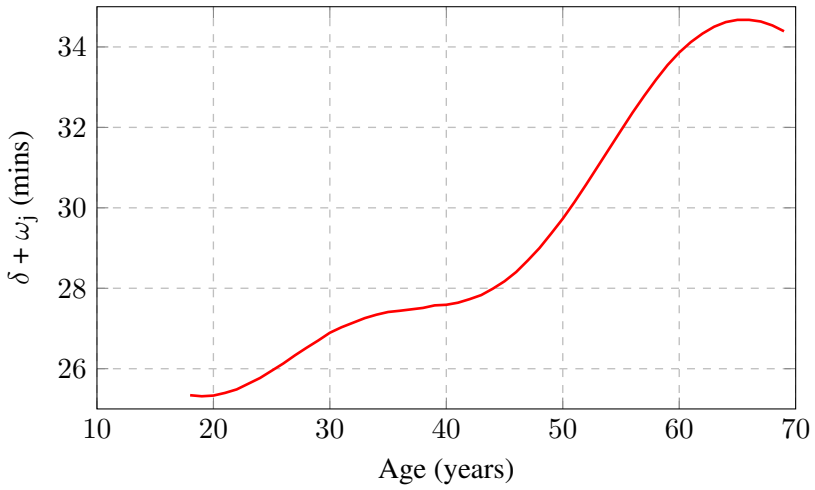


Figure 3.7: **Gender effect on the final performance.** Averaged inferred value for the gender coefficients.

First, we notice that the values of π_{rk} are quite different for each place, but they show little variation between different years. We can argue that this pattern is mainly due to the race difficulty, assuming a stationary selection of the runners.

Boston has the most striking pattern, which can easily be explained by the strict entry requirement time. For 45-49 years old the entry time is 3h25m and the cluster with 70% of the runners has a mean of 3h19m20s. There is a group of almost 15% of the runners that finish just under 4 hours and about 15% of the runners that do much worse than their qualifying time. This might be due to poor training or having some issue during the race. For the Boston marathons, there are no runners in the 3h cluster and around 1% runners in the fastest cluster. The void in the 3h group is due to the massive proportion of runners in the 3h19m group, which makes any runner in that group to be represented by the 3h19m cluster. The runners under 3h are the runners that do much better than the needed qualifying time and cannot be represented by the massive 3h19m group.

The proportions in NYC and London are more similar to each other, as both marathons allow runners to enter the race in more ways than just by entry requirement.[7] The 3h group is more populous in London than NYC, but the 3h19m and 3h53m clusters contain a larger proportion of NYC runners. The 4h30m group is equally probable in both races. London seems to attract a higher proportion of slower runners (over 5 hours). This difference might be due to the difficulty of the marathons (profile and weather conditions) or the pull of runners. NYC race is more hilly than London, which can explain the difference in the first cluster, but the runners in NYC are more diverse (coming from different parts of the country and world), while London attracts more local runners.

[7]Being these two races more accessible or *democratic*, we can consider the proportions in the different clusters closer to the general population of marathon runners.

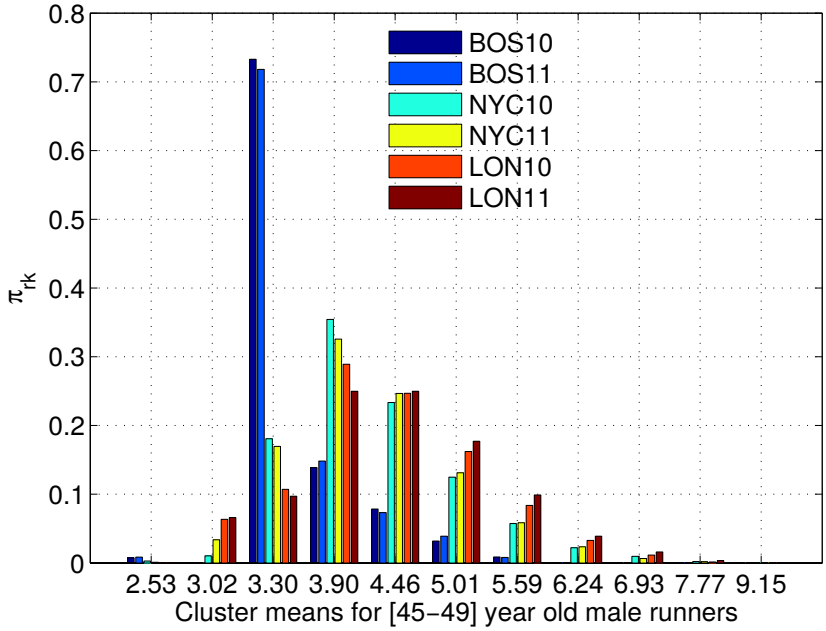


Figure 3.8: **Mixture weights for the H-ADDP mixture model.** The figure shows the mixture weights π_{rk} for each race r and cluster k. The legend shows the different races, and the x-axis corresponds to different clusters.

This might also explain the pattern of the slowest athletes.

3.4.3 Accounting for the Speed of Runners

We now apply the model extension in Section 3.3.2 with cluster-dependent shift delays θ_{jk}. Figure 3.9 and Figure 3.10 show the inferred cluster means for men and women respectively. Although the overall shape of the curves is quite similar across clusters, which validates our previous conclusions, there are some noticeable differences for the fastest runners and we concentrate on those.

The most interesting difference is the behavior of the fastest cluster for under 50 years-old runners, as to the naked eye it seems to suggest that women are faster than men. The fastest cluster for women captures the Olympic runners that are doing under 2h45m, and its proportion is very low (less than 1%). The second fastest cluster for women covers those runners doing under 3h45m and it represents 13% of the female runners. There is a significant difference between Olympic female runners and competitive female runners, so two clusters are needed. For men under 50, the first cluster represents 13% of the runners and it captures those doing under 3h30. The model considers that the Olympic runners can be modeled by the tail of the distribution of competitive runners, without requiring a new cluster as the Olympic women need. Male runners over 50 behave as women do, and two clusters are needed to separate the Olympic and the competitive groups.

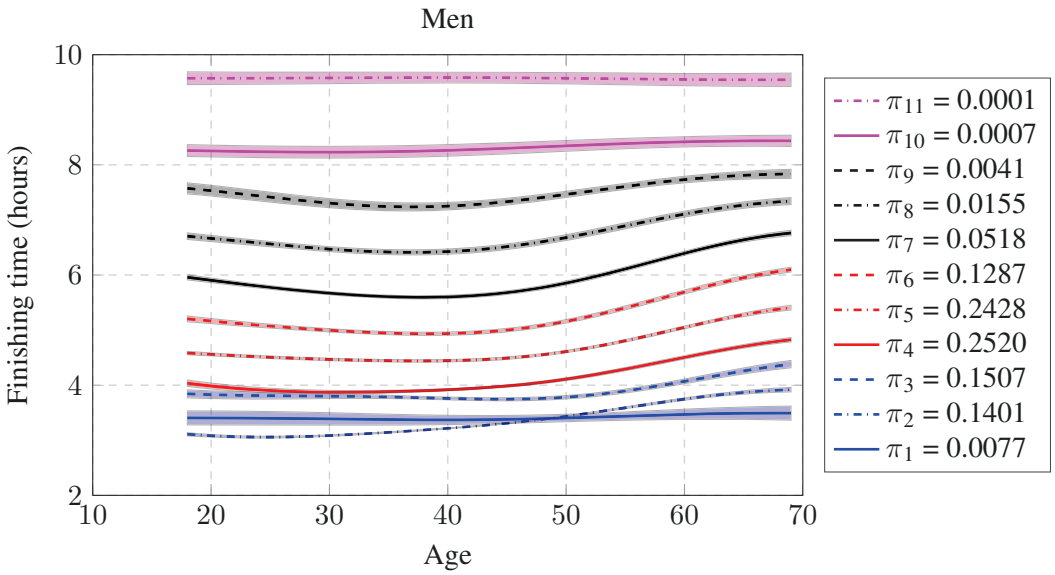


Figure 3.9: **Inferred cluster means $\mu_k+\theta_{jk}$ for men.** We have used the extended model with speed-dependent clusters. The legend shows the inferred value of the proportions π_k for each cluster.

For men under 50, the small cluster that represents the fastest women sits in between the two populous clusters, becoming irrelevant in terms of density estimation. The reason why it appears is because the model forces the same proportion for all clusters across age and gender groups. In order to support this conclusion we have depicted the histogram of the 28-year-old runners together with the inferred density in logarithm scale in Figure 3.11a and Figure 3.11b. In this figure, the blue solid line represents the inferred distribution, the green dash-doted line is the histogram with 6-minute bins, and the red dashed lines represent each one of the clusters multiplied by their averaged weight. In this plot, we can see that the Olympic male runners just doing over 2 hours can be modeled by the same cluster as the competitive male runners, while the finishing time for the Olympic female runners could not be explained by the competitive female ones, and hence a specific cluster is needed. In the previous section, when θ_j was not allowed to vary with k, the cluster for the Olympic males was visible, but this is an effect of forcing the same value of θ_j for all clusters, as women and older male runners need it. This is the only significant difference when we replace θ_j with θ_{jk}.

There may be several explanations for this effect. In the NYC marathon the Olympic women run by themselves in an early wave, while the Olympic men start at the same time as everyone else, so competitive men can try to follow them. However, this does not explain the need for a cluster for fast male runners over 50. We can also hypothesize that female Olympic runners have a training that is significantly different from competitive female runners, while for male runners there is a continuum in the training between Olympic and competitive runners. This could also apply for male over 50,

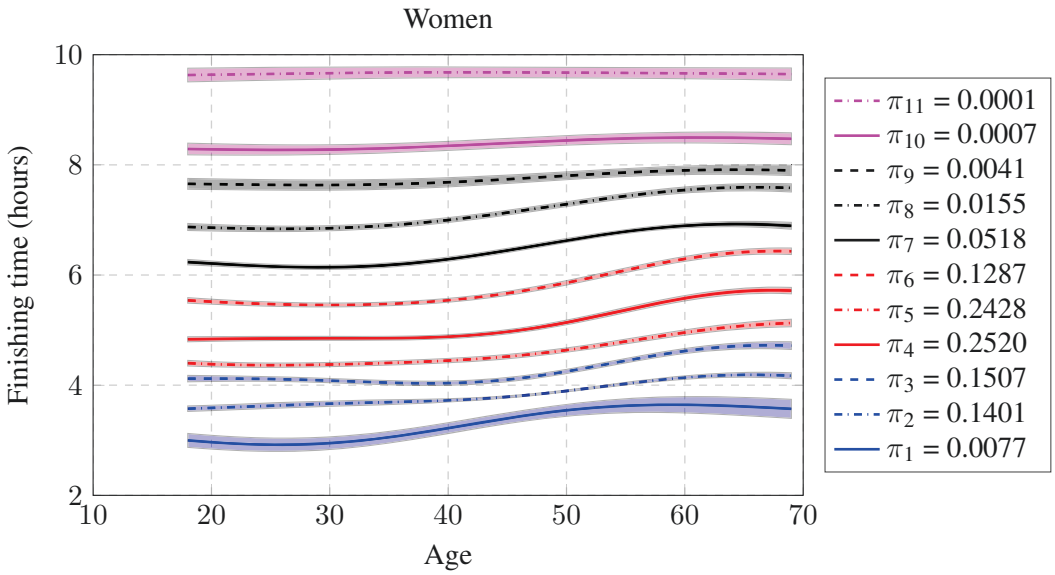


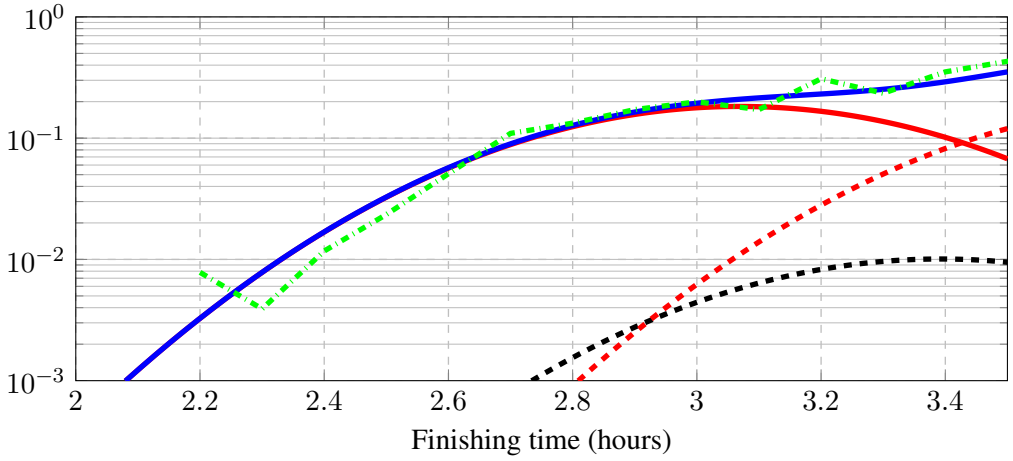
Figure 3.10: **Inferred cluster means $\mu_k + \theta_{jk}$ for women.** We have used the extended model with speed-dependent clusters. The legend shows the inferred value of the proportions π_k for each cluster.

in which there are not that many doing Olympic finishing times and competitive runners are not as strong. Finally, we can also argue that younger male runners are more risky than female and older male runners. Those that succeed do a better time and close the gap between the Olympic and competitive runners. Older males and female competitive runners do not follow such a risky approach and therefore they do not close the gap with the Olympic runners. There is some evidence on this risky hypothesis in Section 3.5.2, in which we see that the reckless running pattern cluster is mainly populated by younger males.

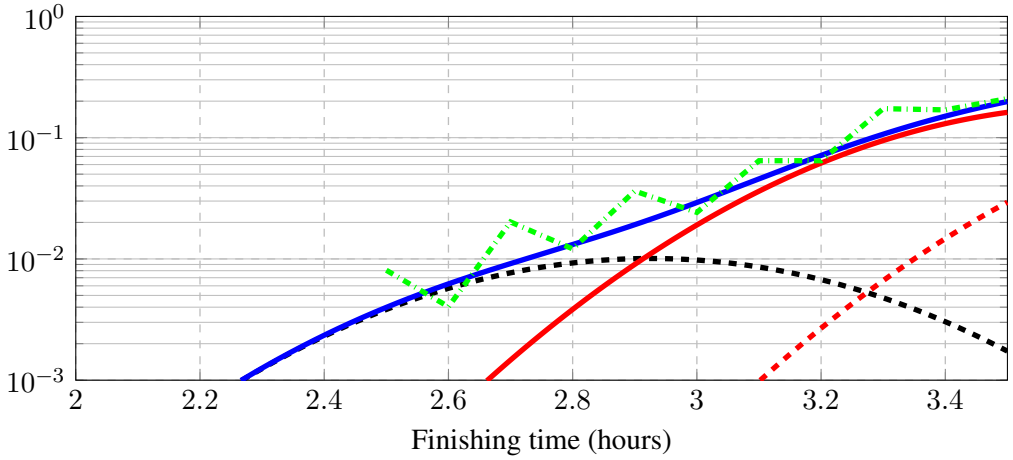
3.5 Analysis of Running Patterns

In our analysis of the marathon dataset, we also led a temporal analysis of the different running patterns. Our objective here is twofold. First, we are interested in a tool that can capture latent running profiles that reflect the marathon difficulties along the 26.2 miles (42.195 km). This can be useful for athletes' training purposes. Second, we aim at predicting the arrival time of runners using intermediate records. This problem has already been addressed in [78], where finishing times are imputed for the 2013 Boston marathon. One of the best approaches rely on the 100 nearest neighbors, which has the limitation of clustering runners that are doing the same absolute times.

We propose to use an HDP [204], a tool that complements the analysis of the previously introduced ADDP, in order to model the fraction of time each runner has spent at each intermediate



(a) Men



(b) Women

Figure 3.11: **Density estimation for 28-year-old runners.** We have used the extended model with speed-dependent clusters. The solid line shows the inferred distribution and the green dash-dotted line is the normalized histogram with 6 minutes bins. Red dashed lines correspond to the individual Gaussian components that define the inferred density, weighted by their proportions. The black dashed line corresponds to Cluster 1.

interval (typically, measures are taken every 5 km and at half-marathon). In this way, we cluster the time ratio instead of the absolute times. Participants that run at different but constant speed will be in the same cluster, no matter if they run each mile in five, eight or eleven minutes. Thus, this model allows estimating finishing times for slower runners that have the same time-ratio profile than fast ones. We use an HDP model in which the likelihood function is a Dirichlet distribution, and each DP clusters the runners by age group and sex. When modeling the full race, it helps to understand the different trade-offs and which parts of the race are harder.

We address the analysis of temporal evolution of runners during the race to understand how the marathoners pace themselves to complete the marathon. We aim at discovering running patterns,

i.e., distinguishing those overly optimistic runners, with decreasing speed along the race, from well-trained runners who tend to keep a constant speed. It also helps to understand where the marathons are harder, so that runners can know beforehand. The hidden running patterns can be used for training purposes, as they can find out the typical shortcomings of athletes with respect to their age, which may help runners train and run more intelligently. In addition, we also show that discovering running patterns provides a new tool for prediction of finishing times, with results comparable to the best reported method in [78].

3.5.1 Modeling

The idea is to cluster the data according to the relative time spent in each interval regardless of each runner's total time. In this sense, we are no longer interested in the absolute times, but in the time proportions invested for each interval. Marathons tend to record the elapsed time every 5 kilometers, in addition to half and full-marathon times. Our input data in this case consists of an $N \times D$ dimensional matrix $\mathbf{X}$ with the time spent for each interval, together with the age and gender. Here, N is the number of runners and D denotes the number of available time records.

We normalize our input data so that each runner is represented as a vector containing the fraction of time spent for each intermediate interval. As in Section 3.3, we split the data $\mathbf{X}$ into J groups of runners having the same age and gender. We use the HDP [204] to cluster the running patterns for the different groups. In the HDP, clusters are allowed to show different probabilities for each group, but the per-cluster parameters are shared across groups. As explained in Section 3.2, we first draw a global base distribution from a DP as $G_0 \sim \text{DP}(\gamma, H)$, where $G_0 = \sum_{k=1}^{\infty} v_k \delta_{\phi_k}$, and for each group j we draw a distribution from a DP using G_0 as base distribution, i.e., $G_j \sim \text{DP}(\alpha, G_0)$. In our model, the likelihood function is a Dirichlet distribution and can be written as

$$\mathbf{x}_{ji} | c_{ji} = k, \mathbf{p}_k \sim \text{Dirichlet}\left(\tau p_{k1}, \ldots, \tau p_{kD}\right), \tag{3.15}$$

where $\mathbf{x}_{ji}$ is the normalized D-dimensional vector for runner i in group j, c_{ji} represents its cluster assignment, $\mathbf{p}_k = [p_{k1}, \ldots, p_{kD}]$ is the vector of patterns representing cluster k, and τ is the concentration hyperparameter of the model. We place a Dirichlet prior over the per-cluster vectors $\mathbf{p}_k$,

$$\mathbf{p}_k \sim \text{Dirichlet}\left(\epsilon \ell_1, \ldots, \epsilon \ell_D\right), \tag{3.16}$$

where ℓ_d is the length of interval d, and ϵ is its concentration hyperparameter.

3.5.2 Experiments

Here, we consider temporal sequences of time measurements every 5 km, and at half and full marathon, as explained in Section 3.5.1. We run $10,000$ iterations of the sampler and average the results for the last $2,000$ iterations.

Running patterns. In this section, we use the data from 2007-2011 NYC marathons, with $194,778$ runners. We discarded data from the 2006 marathon because we observed that intermediate measurements were not fully synchronized with the half and full marathon times. After applying our HDP model, we found a modal value of 46 clusters. In Fig. 3.12 we show the twelve most populated clusters, which account for around 90% of the population on average. The removed cluster do not behave significantly different than the ones we show in this section. For clarity, we do not directly plot the time proportion spent at each interval, but instead show the speed at each 5-km leg, assuming a value of 11 km/h for the speed during the first 5 kilometers. We have removed the half-marathon mark for clarity. The total net time, assuming this value for the initial speed, is shown in the legend for each cluster. We also plot the approximate elevation profile of the marathon with a thin grey line.

Before the half marathon mark, we can roughly see three different types of clusters: those corresponding to athletes running at approximately constant speed (clusters 0, 1, 1^- and 1^{--}), those

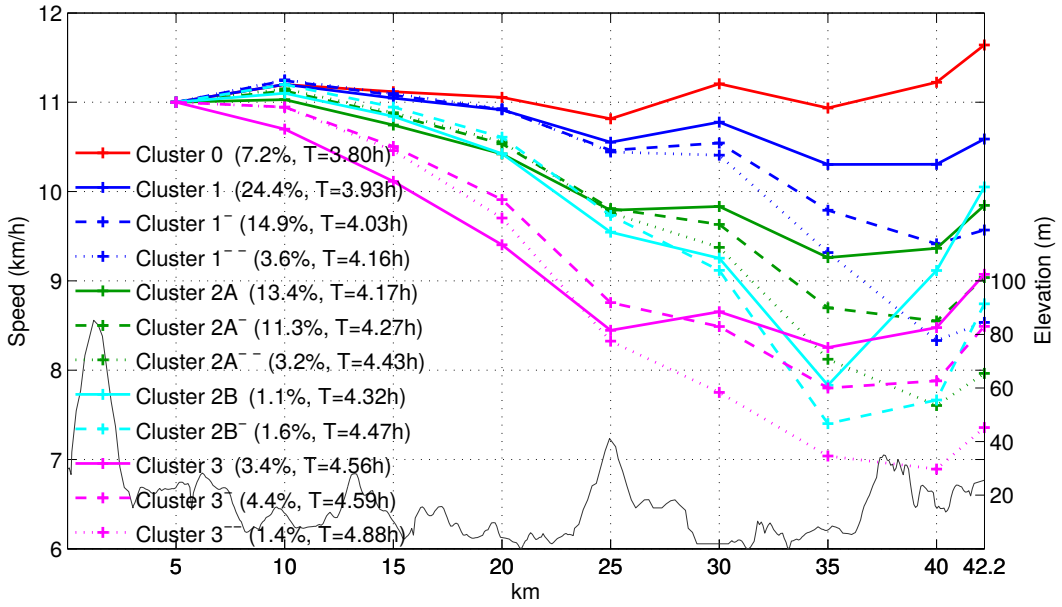


Figure 3.12: **Inferred running patterns by the HDP.** (Thick lines) Inferred running patterns or speed for the twelve most populated clusters, assuming an initial speed of 11 km/h. The legend additionally shows the average proportion of runners in each cluster, as well as the net time for that value of the initial speed. (Thin grey line) Elevation profile of the race.

that are already showing a decreased pace (clusters 2A, $2A^{-}$, $2A^{--}$, 2B, and $2B^{-}$), and those for which the decreased pace is significantly more relevant (clusters 3, 3^{-}, and 3^{--}). Just before the 25-km mark, there is an overall drop in performance that can be explained by the Queensborough bridge and, after that, the twelve clusters become clearly different one another, giving their labels an obvious meaning.

People in Cluster 1 (the most populated cluster, one in every four runner) are well trained runners that run at almost constant speed and the changes can be explained by the hills in each 5-Km interval and they speed up to finish a strong race in the last kilometer, while Clusters 1^{-} and 1^{--} suffered the effect of the Manhattan hills and bridges in and out of the Bronx, besides the natural weariness after running for 35 km. Cluster 1^{--} correspond to the runners that outpaced themselves and finished the marathon at a very low speed, compared with what they could have done. Cluster 0 corresponds to runners who could have done a better race if they had run faster from the beginning and not only after half the race. The rest of the clusters correspond to those overly optimistic runners who could not run as fast as they thought at the beginning. These are the runners who suffered the most. For all the clusters, we can observe an increased speed in the last 2 km, which can be explained by the proximity to the finishing line and the effect of trying to finish under some target time.

Fig. 3.13 shows the averaged inferred proportion of runners π_{jk} in each of the twelve most populated clusters for both men and women, broken down by age groups (blue represents the youngest runners). Clusters 1^{-}, 1^{--}, $2A^{-}$, $2A^{--}$ and $2B^{-}$ are mostly populated by men (e.g., 19.4% of men and only 6.7% of women are in Cluster 1^{-}, and 5% of men and 0.9% of women are in cluster 1^{--}). In other words, the clusters of overconfident runners are mostly populated by men. Clusters $2A^{-}$ and $2A^{--}$ present a constant proportion across ages for both genders. The proportion of women in Clusters 0 and 2A is higher than for men (e.g., 7.5% of men and 24.1% of women are in Cluster 2A). These clusters represent the conservative runners that have some doubts about how fast they can finish a 42.2-km race. In Cluster 0 there is a larger proportion of 18-19 year-old runners for both genders. These are probably first timers, which is consistent with the inexperienced behavior of runners in that cluster. In contrast, Cluster 1 (well-trained athletes) is mostly populated by runners in their thirties, forties and fifties. Custer 2A becomes more popular for older runners. In this cluster, the runner speed slightly decreases in the first part of the race, but it remains constant in the second half-marathon, which might indicate that after the initial 10 km, the runners slow down to make sure that they can continue at a somewhat constant pace.

In Fig. 3.14, we show the averaged proportion of runners in each of the twelve most populated clusters, broken down by their net time, up to 7 hours. As expected, Clusters 0, 1 and 1^{-} comprise

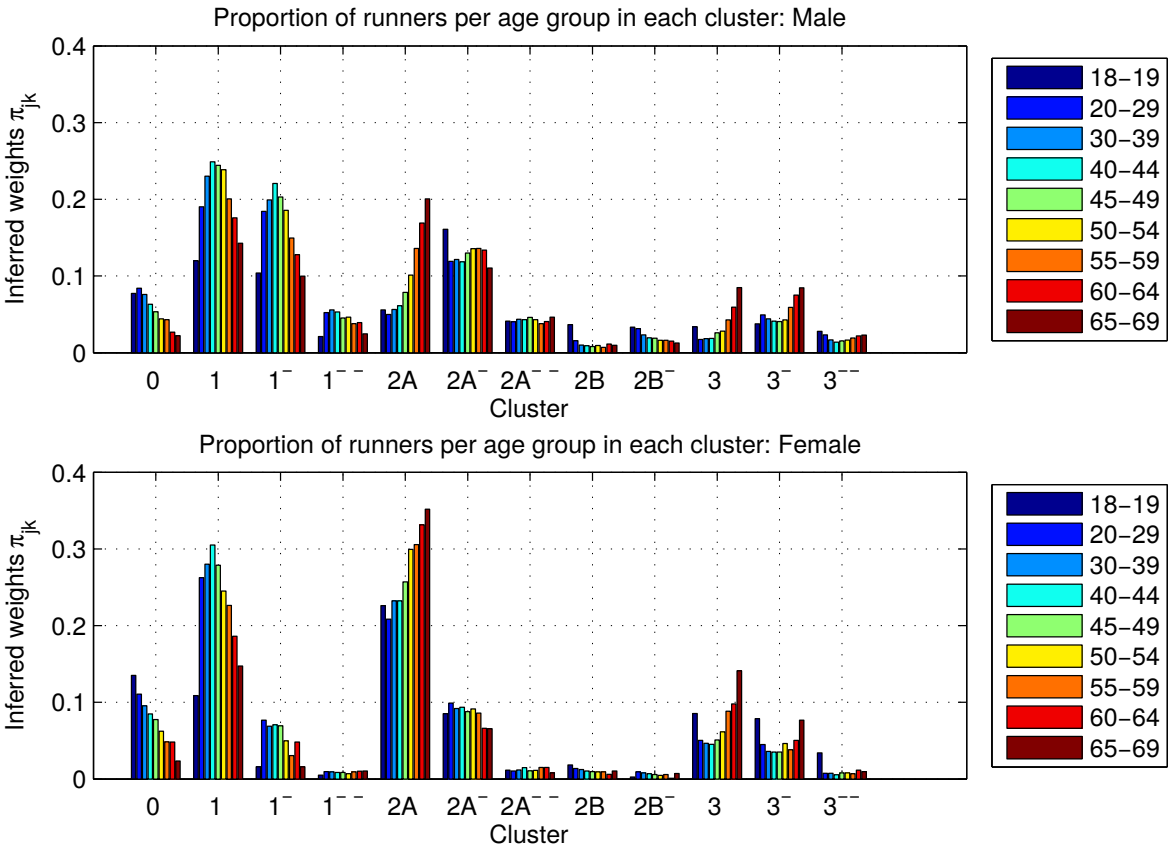


Figure 3.13: **Age proportions for each running pattern.** (Top) Men. (Bottom) Women. The legend shows the different age groups considered.

a high proportion of the fastest runners, i.e., those that can complete the marathon below 4 hours. In contrast, Clusters 3, 3^- and 3^{--} are mostly populated by the slowest runners, with a net time above 5.5 hours. Clusters 2A, $2A^-$, $2A^{--}$, 2B and $2B^-$ have the highest proportion of runners with net time between 4.5 and 5.5 hours. These results are consistent with the description of the clusters provided above.

Prediction of final performance. We can also apply our model to predict the arrival time of athletes. In this case, observations correspond to time proportions at each interval, up to the last available record. We train our model with the subjects in both the test and the training set, assuming that observations up to interval D are known for all of them. Regarding the prediction task, we apply a Bayesian approach in which we take into account the weights from the posterior probabilities of being in each cluster. At each iteration of the sampler and for each runner in the test set, we first compute the posterior probability of being in each of the clusters found using the training set. Second, we project forward his last available time record to obtain the predicted finishing time for each

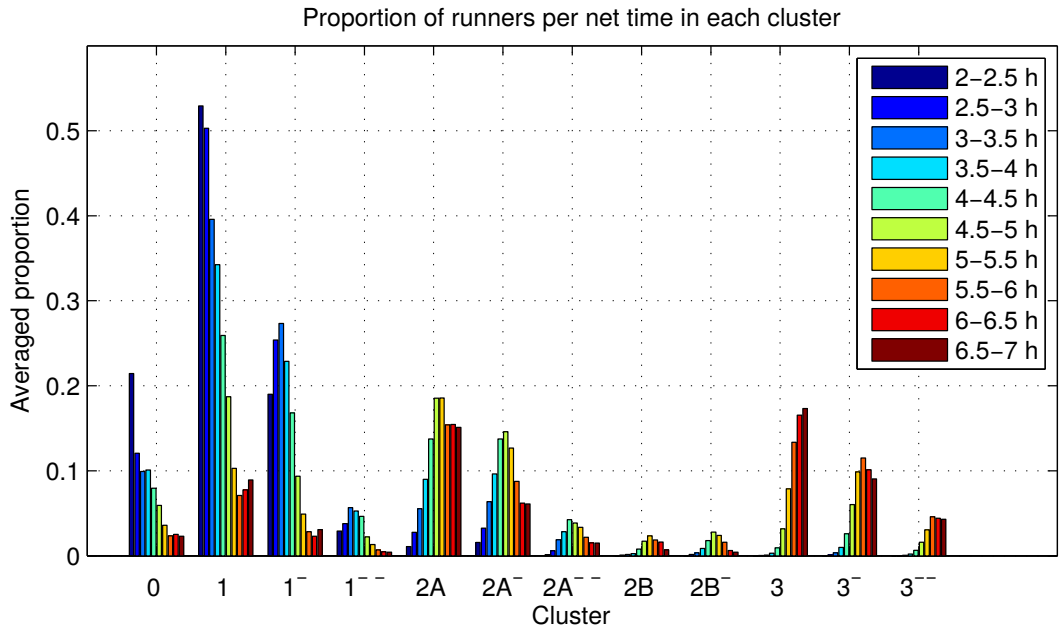


Figure 3.14: **Proportion of runners for each running pattern broken down by net time.** The x-axis indexes the clusters found by the HDP model. The legend shows time intervals for the marathon finishing time.

cluster. Third, our prediction is computed as the weighted average of the predictions for each cluster (weighted by the posterior probabilities of belonging to each cluster). In order to project forward the last available time record for each cluster, we multiply by a factor the time up to interval D of the considered runner. This factor is computed as the median of the quotient of the finishing time and the time up to interval D for those runners in the training set and in the corresponding cluster. Finally, we average our predictions for the last iterations of the sampler.

In Fig. 3.15, we show the empirical density of the prediction error for all subjects in the test set of the 2011 NYC marathon. As the number of available records D increases, the curves tend to shrink around zero.

Table 3.2 reports the average prediction errors, as well as the root of the mean square error, compared with the results obtained following the 100-NN method with forward projection described in [78], for 2010 and 2011 NYC marathons. We do not outperform the discriminative method, but our proposal has the advantage of dealing with time proportions instead of absolute times, which allows predictions for slower runners based on the arrival time of faster ones. Although our model only uses relative times (it has one less degree of freedom) it does equally well, the differences being negligible. Both methods are basically unbiased, as the bias only explains less than 2% of the root of the mean square error, but this bias seems to be always positive, which means that the estimations are optimistic on average.

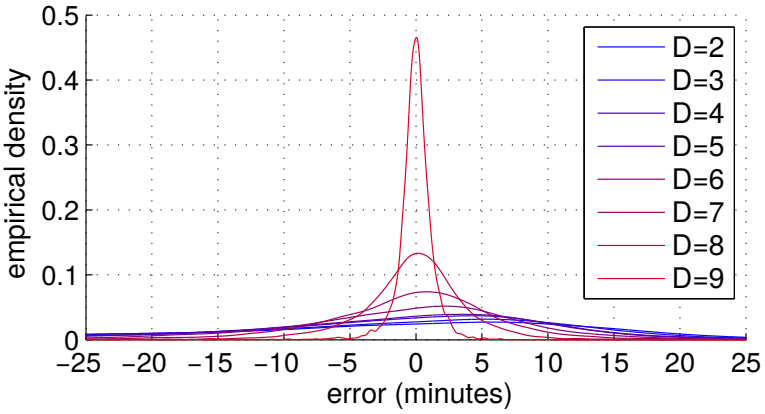


Figure 3.15: **Prediction error.** Density of the prediction error for different values of D (number of available intermediate records).

	100-NN		HDP	
	avg	rmse	avg	rmse
$D = 1$	2.819	19.033	3.778	19.533
$D = 2$	2.554	16.586	3.226	17.224
$D = 3$	1.842	13.408	2.207	14.290
$D = 4$	1.717	12.620	2.000	13.436
$D = 5$	1.264	9.748	1.255	10.536
$D = 6$	0.733	6.913	0.705	7.448
$D = 7$	0.221	3.921	0.212	4.195
$D = 8$	0.031	1.355	0.037	1.434

(a) Predictions for year 2010.

	100-NN		HDP	
	avg	rmse	avg	rmse
$D = 1$	3.384	20.419	4.247	20.104
$D = 2$	2.777	17.124	3.368	17.610
$D = 3$	2.121	13.833	2.428	14.746
$D = 4$	1.942	13.019	2.219	13.788
$D = 5$	1.624	10.202	1.625	10.962
$D = 6$	0.863	7.279	0.821	7.837
$D = 7$	0.283	4.117	0.269	4.440
$D = 8$	0.035	1.394	0.046	1.471

(b) Predictions for year 2011.

Table 3.2: **Test prediction errors for year 2010.** We show the average error for both 100-NN and HDP methods ("avg"), as well as the square root of the mean square error ("rmse"). Results are all expressed in minutes. Rows represent number of available time records.

3.6 Connexion to Infinite Mixture of Experts

The presented ADDP model described in Section 3.3 has been applied so far in this chapter to conduct a data exploratory analysis of marathon races. However, this model can also been used as a general non-linear regression approach, able to handle heteroscedastic noise and arbitrary output likelihoods. Instead of using a single GP that tracks the mean underlying function, we have several GPs that model the underlying distribution for each input vector as an infinite mixture of Gaussians. These GPs cover the whole input space, i.e., globally, allowing the predicted posterior probability to be non-Gaussian, multimodal, heteroscedastic and/or non-stationary, without the need of explicitly indicating or even knowing that those effects might be into play. Our approach is also able to provide accurate percentile information, and can easily be used for Bayesian integral computation. An infinite mixture of global Gaussian processes (IMoGGP) can potentially capture any complex functional behavior, in a similar

fashion that an infinite mixture of Gaussians can approximate any arbitrary density function.

For test input vector, our model provides an estimate of the output that is a linear combination of GPs. The proposed method is a discriminative regression algorithm, in the same way standard GPs are. Hence, we make no probabilistic assumptions over the input space. If we are strictly interested in making a probabilistic statement over the output space given the input, adding a probabilistic model over the input might be detrimental, both in terms of computational complexity and accuracy of our predictions. The mixing proportions are held constant throughout the input space, so that all GPs are active in the whole input domain. Hence, we avoid the need of relying on a gating function, which is typically used to select a particular local functions. The mixing proportions, as well as the hyper-parameters of the Gaussian processes, are inferred given the data. In this inference phase, as we change the mixing proportions and the inputs are shuffled between the different global GPs, we are able to capture any of the effects previously described (e.g. non-Gaussian likelihoods, heteroscedasticity, colored noise or multimodality) without needing to know if they are present.

Our algorithm is universal for solving any regression problem, but it is more effective for moderate-sized input spaces and a larger number of samples per input dimension, because if the input dimension is large the fitting with one Gaussian process will prevail (i.e., we get the standard GP fitting) and if the input dimension were low (or nonexistent), the solution would be that of a DP for a Gaussian mixture. In short, our algorithm transitions seamlessly between the two limiting processes.

In Figure 3.16 a), we show a one-dimensional cartoon solution that our discriminative regression algorithm would be able to provide. Our algorithm is an infinite mixture of experts (IMoE), but not a typical one in which the input space is chopped locally, using a gating function that decides who is the expert for each input (see cartoon in Figure 3.16 b). The proposed approach divides the available data in independent GPs, so it presents the same computational savings as the standard mixture of expert algorithms based on GPs. The data division is not based on local proximity rules, but on improving the prediction accuracy of the output.

Our algorithm is a direct application of the DDP [125]. But our interpretation as a nonparametric universal regression discriminative procedure is novel, as the standard interpretation of DDPs is that of an indexed collection of distributions.[8] Moreover, the DDP should be understood as a general framework instead of a specific algorithm, as many existing models can be explained as such. In Section 3.6.1, we review the literature thoroughly, because our work is related to several well-known algorithms. But we want to emphasize that although our algorithm is similar to others, those have been presented in the past under narrow conditions or specific applications, which do not show the

[8] Actually, the author in [125] proposed a simple 1-D linear regression application in which the strength of the DDP for nonparametric regression and its many desirable properties are not exploited nor hinted.

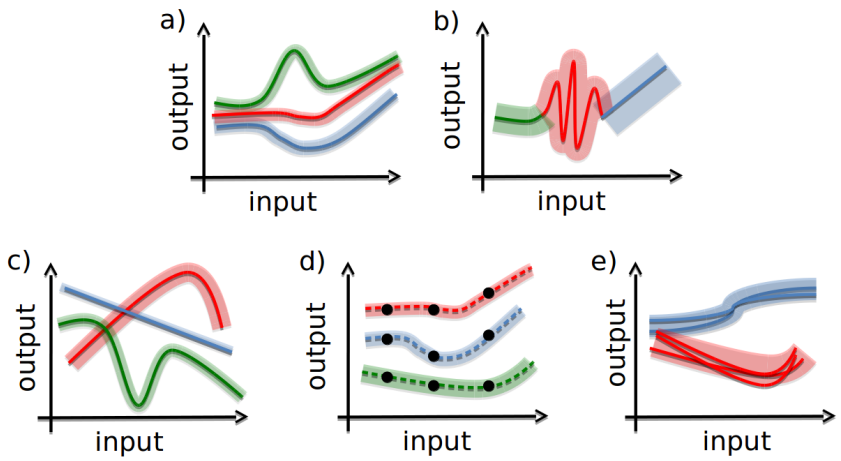


Figure 3.16: **Conceptual comparison of different approaches.** Sketch comparing a) IMoGGP (proposed approach), b) IMoE, c) Overlapping GPs for multi-tracking, d) Spatial DP, and e) time series clustering. Each color represents a different GP.

potential of such algorithm for general regression with the properties described earlier.

For example, in [123], the authors infer trajectories, i.e., time series, in which each trajectory is represented by a GP, as shown in Figure 3.16 c), but there is no regression interpretation nor future predictions, and the number of time series has to be known beforehand. In [58], the authors want to estimate the 10-day aggregated rainfall in 39 locations in southern France and 6 test locations. They assume that the input space has a low cardinality of potential locations, and each sample had an observation in all input locations. As sketched in Figure 3.16 d), the output is a collection of vectors, each one with same dimension as the cardinality of the input space; this is far from a typical regression scenario and can be seen as a particular case of our method. Our algorithm is also similar to the clustering of time series using DPs [86], as illustrated in Figure 3.16 e), but in that case samples are treated as whole sequences a priori, and the goal is to identify clusters. Our algorithm would be able to solve these applications directly or with minor modifications, while the solution in those papers cannot be applied to solve the general regression problem.

3.6.1 Related Works

In this subsection, we review the literature that is related to our approach and indicate the main differences with our proposal. We have carried out a thorough literature search and, for brevity, we only reference those that are directly related to our approach, so many papers that are not directly comparable have been left out. We know this is a widely-research field and if any relevant body of work has not been referenced, it has been unintentional.

Modeling with Mixture of GP Experts

Mixture of expert models in which each expert is a GP has been proposed in the literature in several occasions. One of the most popular methods in Bayesian regression is the IMoE to capture local properties of the signal in different areas of the input space [63]. A gating function is used to determine which GP is active in each area, making it useful locally. Extensions of the IMoE include modeling the input and output jointly, i.e., modeling $p(\boldsymbol{x}, y)$ [128], or allowing the use of the same experts in different input areas through hierarchical DPs [199]. Our work addresses GP components at a global scale, which is a more natural way to capture heteroscedastic noise and other gradual behaviors. Global GP priors were previously proposed in [113], but only considering a fixed number of mixtures. Also the work in [195] proposes a mixture of global Gaussian Processes for traffic flow prediction, but their approach is generative and models the input space too.

Multiresolution GPs [55] relies on a hierarchy of GPs to partition the whole space in order to capture long-range, non-Markovian dependencies while allowing for abrupt changes. Our method does not partition the input space and allows for multiple functions at a same input location instead. Additive GPs in [48, 162] divide the output function into low-dimensional components of varying degrees. Our method is different as we allow for multiple output functions instead of a partitioning of the output dimensions, i.e.

$$g(y_i) = \sum_{k=1}^{\infty} f_k(x_{i1}, \ldots, x_{iD}) \qquad \text{(IMoGGP)}$$

$$g(y_i) = f(x_{i1}) + f(x_{i2}) + f(x_{i3}, x_{i4}) + \ldots \qquad \text{(aGP)}$$

The additive GPs approach seems more suitable for high-dimensional input spaces, it fails to capture heteroscedasticity or multimodality, as our proposal does.

Density Regression with Dependent Dirichlet Processes

DDPs are useful to model collections of distributions that vary in time, space or experimental settings. In the literature, it is often the case to use a semiparametric model, i.e., a parametric function for the signal, and a nonparametric prior for the noise, in order to capture heteroscedasticity or non-Gaussianity [47, 178, 58, 44, 71]. Our model considers a more general formulation by assuming a completely nonparametric model for both signal and noise.

In [178], a DDP prior is used to model the joint distribution $p(\boldsymbol{x}, y)$, given different experimental conditions. With such a generative approach, modeling $\boldsymbol{x}$ might dominate over y, resulting in an under-fitting of y. Our approach directly focus on the conditional distribution $p(y|\boldsymbol{x})$, and applies

the DDPs in a different way, directly over the input space $\boldsymbol{x}$. Such discriminative perspective typically gives better accuracy and has the advantage of estimating less parameters than in the generative approach. The work in [37] uses sp-DDPs, i.e., DDPs with constant weights over $\boldsymbol{x}$, to cluster the behavior of multiple ANOVA models under different experimental conditions. Here again, their approach is generative, whereas we use the DDP to directly model the conditional distribution $p(y|\boldsymbol{x})$.

In [47], DDPs are used for Bayesian density regression with kernel-varying weights, assuming a linear relationship between y and $\boldsymbol{x}$. Our approach generalizes this work by replacing the linear basis functions by arbitrary non-linear functions with a GP prior each. The authors in [58, 44] introduce a DDP-based model for spatial modeling applications called the spatial DP prior, which is a probability weighted collection of random surfaces. They use a linear process for the signal and a mixture of GPs to capture the noise. Because each atom in the DP corresponds to a realization of a random field over the input space, their algorithm needs the assumption that multiple points are available at each location $\boldsymbol{x}$ and in particular, that the number of points assigned to each GP is always uniform. Our approach removes those restrictions and assumes no particular functional form of the signal.

Clustering of Time Series

Finally, hierarchical mixtures of GPs have often been used in the literature to cluster time series, as in [190], [86] and [180]. All these works assume prior knowledge of which points belong together to the same temporal sequence, and assign the points of a temporal sequence jointly to the same GP. All these models are generative and seek interpretable results. In our case, cluster assignments are purely auxiliary variables, we only care about predictive accuracy. Slightly different and also closely related to our approach is the work in [123], which uses a parametric mixture of GPs for the data association problem. The objective there is to find the appropriate cluster assignments for each point and recover multiple trajectories, which is useful in multi-tracking scenarios. In our case, mixture assignments are just auxiliary variables, and we assume a potentially infinite number of mixtures to represent the data.

3.6.2 Infinite Mixture of Global Gaussian Processes

The aim in a regression problem is to estimate $y \in \mathbb{R}$ given an input $\boldsymbol{x} \in \mathbb{R}^D$ and a database $\mathcal{D}_n = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$. From the available data it induces a general relation between the input $\boldsymbol{x}$ and the output y. In probabilistic modeling this relation is expressed by a conditional model:

$$p(y|\boldsymbol{x}, \mathcal{D}_n). \tag{3.17}$$

To introduce our IMoGGP we are going to start from the standard stick-breaking construction of DPs for countably infinite mixture models [200]. Observations are then generated as follows:

$$\boldsymbol{\pi}|\alpha \sim \mathrm{GEM}(\alpha) \tag{3.18}$$

$$z_i|\boldsymbol{\pi} \sim \mathrm{Multinomial}(\boldsymbol{\pi}) \tag{3.19}$$

$$\theta_m|H \sim H \tag{3.20}$$

$$y_i|z_i, \{\theta_m\} \sim F(\theta_{z_i}), \tag{3.21}$$

where GEM stands for the stick-breaking prior by Griffiths, Engen and McCloskey [152], α is the concentration parameter of the DP, and the mixing proportions $\boldsymbol{\pi}$ are sampled using a stick breaking procedure [99]. z_i indicates the cluster assignment of observation i, and θ_m designates the cluster parameters for cluster m, which are sampled from the base measure H. Finally, observations are sampled from $F(\cdot)$ given the cluster assignments and parameters. One standard selection for $F(\cdot)$ is a Gaussian distribution in which θ_m represents its mean and variance.

In the regression setting, each y_i is associated with an input $\boldsymbol{x}_i$, so (3.21) can be modified as

$$y_i|z_i, \{\theta_m\}, \boldsymbol{x}_i \sim F\big(\theta_{z_i}(\boldsymbol{x}_i)\big), \tag{3.22}$$

$$\theta_m|H, \phi_m \sim H_{\phi_m}, \tag{3.23}$$

where we assume that $F\big(\theta_{z_i}(\boldsymbol{x}_i)\big)$ is Gaussian-distributed with mean $\mu_{z_i}(\boldsymbol{x}_i)$ and variance $\sigma^2_{z_i}(\boldsymbol{x}_i)$, and H_{ϕ_m} is a Gaussian process prior with hyperparameters ϕ_m. Now each cluster parameter θ_m corresponds to a latent function over the input space. This construction with a general $F(\cdot)$ is exactly a sp-DDP where the mixture weights are constant over the input space [125], and the parameters of each component of the infinite mixture model is indexed by the input variable $\boldsymbol{x}$.

The inference in this model is straightforward, because given $\{\theta_m\}$, sampling z_i and $\boldsymbol{\pi}$ is identical to the inference of cluster assignments in DPs. Given the number and proportions of clusters and cluster assignments for all pairs $(\boldsymbol{x}_i, y_i)$, $\{\theta_m\}$ can be inferred by sampling from the posterior GP distribution, and ϕ_m can be obtained by sampling (or maximizing) the evidence for each individual GP in parallel [172]. Specifically, we perform inference by a simple Markov chain Monte Carlo (MCMC) procedure. We first use the auxiliary variable approach from Algorithm 8 in [140], i.e., we do not integrate out the hidden variables, and we propose T new clusters at each iteration, which then allows to sample z_i in parallel. Hyperparameters for potential new GPs are sampled from the prior at each iteration. The procedure for each iteration is detailed in Algorithm 1, in which we have used the standard vector notation, i.e. $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top$ and $\boldsymbol{z} = [z_1, \ldots, z_n]^\top$. To

evaluate the GP posterior, i.e., $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}, \{\theta_m\})$, we only need to consider the subset of points that are related to each GP, providing huge computational gains.

Algorithm 1 Inference for the IMoGGP

I - Initialize all hidden structures.

II - For each iteration of the Gibbs sampling procedure:

1: Sample extended vector of mixture proportions:

$$\boldsymbol{\pi}|\boldsymbol{z}, \alpha \sim \text{Dirichlet}\big(n_1, \dots n_K, \underbrace{\alpha/T \dots \alpha/T}_{T \text{ times}}\big) \tag{3.24}$$

2: Sample latent functions, i.e., cluster parameters θ_m, $m = 1, \dots, M^+$:

$$p(\theta_m|\boldsymbol{\pi}, \boldsymbol{y}, \mathbf{X}, \boldsymbol{z}) \propto\ p(\theta_m|H_{\phi_m})p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}, \theta_m) \tag{3.25}$$

3: Sample cluster assignments:

$$p(z_i|\boldsymbol{\pi}, y_i, \boldsymbol{x}_i, \boldsymbol{z}_{-i}, \{\theta_m\}) \propto p(z_i|\boldsymbol{\pi})\ p(y_i|\boldsymbol{x}_i, \boldsymbol{z}, \{\theta_m\}) \tag{3.26}$$

4: Sample hyperparameters ϕ_m, $m = 1, \dots, M^+$:

$$p(\phi_m|\boldsymbol{\pi}, \boldsymbol{y}, \mathbf{X}, \boldsymbol{z}) \propto\ p(\phi_m|H)p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}, \phi_m) \tag{3.27}$$

5: Sample concentration parameter α for the mixture model

The prior probability of assigning a data point to a new GP is proportional to the concentration parameter α. This hyperparameter directly influences the total number of GPs used to model the data. We sample α using an auxiliary variable η as in [50]. Allowing α to vary makes the model more flexible and ready to be used for different datasets.

Finally the predicted distribution for a new input $\boldsymbol{x}^*$ is given by:

$$p(y^*|\boldsymbol{x}^*, \mathcal{D}_n) = \sum_{m=1}^{M^+} \pi_m p(y^*|\boldsymbol{x}^*, \mathcal{D}_n, \boldsymbol{z}, \{\theta_m\}), \tag{3.28}$$

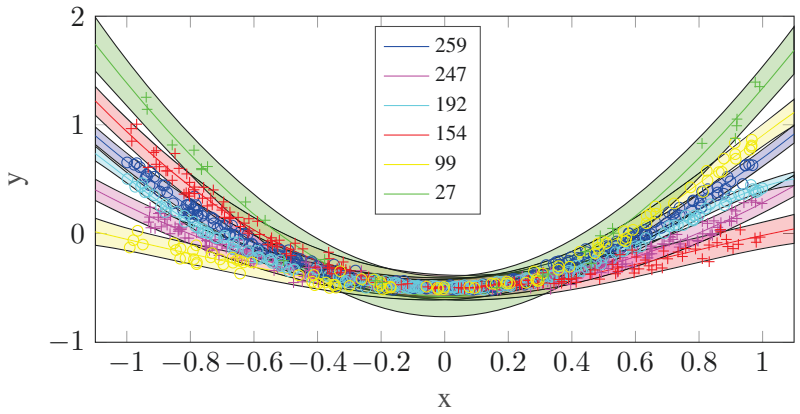
where

$$p(y^*|\boldsymbol{x}^*, \mathcal{D}_n, \boldsymbol{z}, \{\theta_m\}) = \mathcal{N}(\mu_m(\boldsymbol{x}^*), \sigma_m^2(\boldsymbol{x}^*)), \tag{3.29}$$

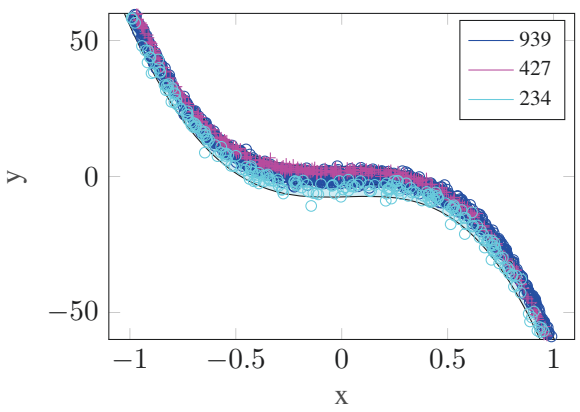
$$\mu_m(\boldsymbol{x}^*) = \mathbf{k}_m^\top \mathbf{C}_m^{-1} \boldsymbol{y}_m, \tag{3.30}$$

$$\sigma_m^2(\boldsymbol{x}^*) = k_m(\boldsymbol{x}, \boldsymbol{x}) - \mathbf{k}_m^\top \mathbf{C}_m^{-1} \mathbf{k}_m, \tag{3.31}$$

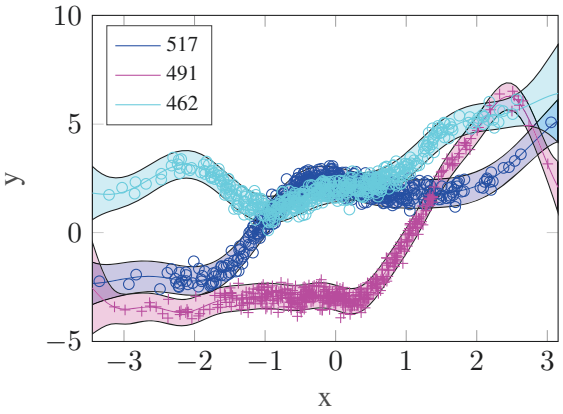
and $\mathbf{k}_m = [k_m(\boldsymbol{x}_1^m, \boldsymbol{x}^*), k_m(\boldsymbol{x}_2^m, \boldsymbol{x}^*), \ldots, k_m(\boldsymbol{x}_{n_m}^m, \boldsymbol{x}^*)]^\top$, $\mathbf{C}_m = \mathbf{K}_m + \sigma_m^2 \mathbf{I}$. The set $\{\boldsymbol{x}_j^m\}_{j=1}^{n_m}$ are the $\boldsymbol{x}_i$ for which z_i is equal to m and $(\mathbf{K}_m)_{rs} = k_m(\boldsymbol{x}_r^m, \boldsymbol{x}_s^m)$. Finally $k_m(\boldsymbol{x}, \boldsymbol{x}')$ is the kernel or covariance function for each GP. This covariance function influences the behavior of each individual GP, and popular choices are a squared exponential or Mattern kernels, a positive linear mixture of kernels or their products, see [172] for further details about valid covariance functions.



(a) Heteroscedastic Noise.



(b) Non-Gaussian Likelihoods.



(c) Multimodal Distributions.

Figure 3.17: **Properties that can be captured by the IMoGGP model**: (a) non-stationary, heteroscedastic noise; (b) non-Gaussian likelihoods, specifically a Student's t with Gamma distributed noise; and, (c) multi-modal predictive distributions. Point assignments to different GPs are represented with different colors and shapes. The legend lists number of points associated to each GP.

In Figure 3.17, we illustrate with synthetic data the complex behaviors and interesting properties that our IMoGGP is able to capture. Figure 3.17a shows a quadratic function $y = x^2 - 0.5 + \epsilon$ with added input-dependent Gaussian noise $\epsilon \sim \mathcal{N}\big(0, (0.01 + \sin(2\pi x/10)^2)^2\big)$. In Figure 3.17b, we show a cubic function $y = 4x^3 - 1 + \frac{1}{2}\epsilon + 3\gamma$ with added Student's t and gamma noise, $\epsilon \sim \text{Students}'\ \text{t}(10)$

and $\gamma \sim \text{Gamma}(2, 0.5)$.[9] In Figure 3.17c, we have generated 3 GPs with a Mattern covariance functions plus Gaussian noise. In all these examples the underlying GP covariance function was a squared exponential and the likelihood model was Gaussian. It can be seen that our IMoGGP model is able to deal with heteroscedastic data, heavy tailed and asymmetric noise, and parallel Gaussian processes with the same regression model, by adding global GPs over the whole input space to capture these behaviors.

3.6.3 Simulations

This Section compares the performance of the proposed model IMoGGP against a single GP and the IMoE from [171]. The three algorithms are compared in exactly the same conditions, with the same hyperparameters and input splits into training and test data. Each simulation is run for 1000 iterations, and averaging is done for the last 500 iterations. Because data points are clustered in multiple GPs, our algorithm benefits from the same computational gains than in the IMoE case, and both algorithms are much faster to train than a single GP. All results are computed on an independent test set corresponding to 20% of the total input data, and we perform 10 different splits at each time. For all our experiments, we use the popular Noisy Squared Exponential (NSE) kernel, given by

$$K(\boldsymbol{x}, \boldsymbol{x'}) = v \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x'}\|^2}{2w_d^2} \right) + v_0 \delta(\boldsymbol{x}, \boldsymbol{x'}) \tag{3.32}$$

where v is the signal variance affecting the output range, v_0 is the noise variance and w_d is the length scale per dimension d of $\boldsymbol{x}$, which controls how quickly the function can vary, e.g. bigger values for w_d result in a smoother function for dimension d. This kernel has been proved to be universal, i.e., a single GP with a noisy squared exponential kernel can approximate arbitrary smooth functions if given enough data [130]. For both the IMoGGP and IMoE models, the overall computation is decreased by separating data points in different GPs. For scalability reasons and fairness, we compare our algorithm against an uncollapsed version of the IMoE algorithm presented in [171].

Database descriptions. The synthetic databases correspond to the examples in Section 3.6.2, and include $n = 2000$ observations for each case. We also consider three different real databases, all of them publicly available. The concrete database from [223] consists of 1030 observations and the input dimension is 8. The objective is to predict the compressive strength (MPa) of concrete, which is one of the most important materials in civil engineering. This is a highly nonlinear function of age and ingredients that include cement, blast furnace slag, fly ash, water, superplasticizer, coarse

[9] We define the Gamma distribution in terms of shape and rate parameters.

		Heteroscedastic	Non-Gaussian	Multimodal	Concrete	Marathon	RSSI
LLH	**GP**	-0.0217	-3.4920	-3.3030	-0.5855	-1.6373	0.2033
	IMoE	0.7017	-2.1248	-2.1604	1.9452	-1.6308	**0.9943**
	IMoGGP	**0.9008**	**-2.1237**	**-1.2575**	**2.3587**	**-1.5723**	0.9846
MSE	**GP**	0.0288	**4.8115**	4.2815	93.4640	0.7877	86.6815
	IMoE	0.0331	4.8394	5.2263	93.4640	0.7780	82.8929
	IMoGGP	**0.0287**	4.8500	**4.2703**	**43.6710**	**0.7754**	**82.4264**

Table 3.3: **Comparison of Accuracy.** Average test LLH and MSE for the single GP, the IMoE and the proposed IMoGGP. The three first databases correspond to the toy examples plotted in Section 3.6.2. The last three columns correspond to real databases described in this section and publicly available online.

aggregate, and fine aggregate. The actual concrete compressive strength (MPa) for a given mixture under a specific age (days) was determined from laboratory.

For the New York City marathon data, the objective is to predict the arrival time of a runner given his gender and age. The output y designates the arrival time, and $\boldsymbol{x}$ is a two dimensional vector with the age and gender for each runner. We take a subset of $4,800$ runners in total, keeping the same age/gender relative distribution.

Finally, the RSSI database consists of 4799 measurements of the Received Signal Strength Indicator (RSSI), which captures the power of different wireless networks at different locations along a large corridor.[10] Modeling RSSI correctly is very important, as different signal strengths can have a strong impact on functionality in wireless planning and localization [15].

Results. Table 3.6.3 shows quantitative results for both synthetic and real databases. We report the mean predictive log likelihood (LLH) and mean square error (MSE) for each method and input database. Bold values designate the best method in each case. In this experiments the PLLH is more telling of the quality of the achieved results, because it captures the full output density estimate, while the mean only measures the quality of the mean prediction. Nonetheless, we also report the mean for completeness. For the same experimental conditions, our method gives the highest LLH in five out of six databases.

The highest gains are achieved for the Multimodal and Concrete datasets. Indeed, the IMoGGP is the only method able to use multiple functions at a single input location. On the other hand, the Concrete database is the example with highest dimension ($D = 8$), and the curse of dimensionality makes it harder for the IMoE to learn local functions. The IMoGGP is less affected as it uses global GPs over the input space and is able to share more information across all dimensions.

The presented approach is simple, yet powerful to solve regression problems by using a IMoGGP

[10]This database is available in my personal webpage.

that cover the whole input space. This method can deal with arbitrary output distributions, non-stationary signals, heteroscedastic noise and multimodal predictive distributions straightforwardly. It is suitable for very large databases of moderate dimension. The computational cost is not enormous as the data points are partitioned in several GPs, and the sampling step of cluster assignments can be run in parallel. As future work, we would like to extend this work to deal with high input dimensions. In such scenario, it might be desirable to have varying mixture weights across the input space like in [30] and a selection of relevant input dimensions and orders, such as in [48].

3.7 Summary

In this chapter, we have presented a novel application of BNP to model marathon runners. By including constraints over a sp-DDP prior, we have provided insightful solutions to the problem of age-grading in marathon races. The ADDP model informs us of the impact of age and gender on runners' performance. Statistical age-grading curves have been inferred, which allow us to rank and reward athletes in a fair manner in official events. A model extension lets us capture the impact of meteorological conditions or even topology on runners. We infer the latent difficulty for each competition at different years, allowing robust comparison across different marathon events.

Regarding the applicability of our approach, we remark that the idea of comparing group density distributions fairly within a single model is an attractive research path, that could result in a huge and broad number of applications. It can be applied to many problems involving stratified data and a certain control variable (e.g., age, gender, nationality). In problems concerning group data or any competitive human activity, sharing the mixture weights across groups is a sensible assumption. Some application examples can be found in pediatrics (e.g., comparison of children population according to weight and height), social sciences (e.g., analysis of gender impact on actual salary income across countries), or pharmaceutics (e.g., monitoring certain drug responses according to some patient covariates).

The novelty of the proposed approach relies not only on the application, but also on the necessary steps to transform a prior that provides accurate estimates into a prior that also gives interpretable results. Non-trivial structural assumptions and design solutions are made to find hidden properties of the athletes while providing accurate predictions. We believe that BNP models will be more useful for non-machine-learning experts in the future if we can tailor the priors to provide accurate and understandable solutions.

4

Case-control Indian Buffet Process for Analysis of Clinical Trials

Biomarker discovery in clinical trials is of the utmost interest to understand both disease mechanisms and drug effects on patients [136]. Traditional approaches for screening potentially interesting biomarkers suffer from at least one of the following: *population heterogeneity*, i.e., the fact that each individual has unique characteristics at all levels (demographics, environment or biological aspects) resulting in different disease progressions and drug responses; *strong correlations* among biomarkers that might difficult their detection if screened individually; *difficulty to isolate drug effects* from natural patient response, specially in small sample-size scenarios.

This chapter presents a novel Bayesian nonparametric (BNP) method for subpopulation characterization and biomarker discovery in clinical trials which overcomes the aforementioned problems. The case-control Indian buffet process (C-IBP) allows for the identification of prognostic and predictive variables globally and specifically to each subgroup. It extends the general latent feature model (GLFM) [210] by additionally sharing information across case-control patients in a structured way. Empirical results on a phase II clinical trial for liver cancer demonstrate that the C-IBP can find already well-known relevant biomarkers and discover statistically significant new ones.

4.1 Introduction

Clinical trials constitute a key research tool for advancing medical knowledge and patient care [103, 179]. They are crucial not only in the assessment of drug efficiency and undesired side-effects, but also in the understanding of complex biological mechanisms, and discovery of interesting biomarkers for personalized diagnosis and precise treatment [85, 100]. In this chapter, we refer to *biomarker* as any variable that can be used as an indicator of a particular disease state. Biomarker discovery is central in the study of disease mechanisms, and is particularly useful for prediction of disease progression, prescription of appropriated drugs based on patient characteristics, or even as potential targets in the development of new drugs, tailored to specific subpopulations [32].

Ultimately, we seek to identify two types of biomarkers: *prognostic* and *predictive* variables. A prognostic factor is a clinical or biological characteristic that provides information on the likely outcome of a patient disease without considering any treatment. In contrast, a predictive factor provides information on the likely benefit from treatment, e.g., in terms of tumor shrinkage or survival [136, 92]. Such predictive factors can be used to identify subpopulations of patients who are most likely to benefit from a given therapy. Some good examples of predictive biomarkers being used in the daily clinical oncology practice are estrogen and progesterone receptors to predict sensitivity to endocrine therapy in breast cancer, HER2 to predict sensitivity to Herceptin treatment, and KRAS mutation to predict resistance to EGFr antibody therapy [85]. Both types of biomarkers are typically used in precision oncology for molecular diagnosis of chronic myeloid leukemia, colon, breast, and lung cancer, as well as in melanoma [139].

Looking for interesting biomarkers has proved so far to be a very challenging task given the high number of potential candidates, inherent complexity of biological mechanisms, and patient heterogeneity [108]. Indeed, clinicians might have a certain amount of information about their patients, including genetic data, clinical observations, lab measurements, or environmental factors. Data are typically quite heterogeneous, noisy (might contain missing data), high-dimensional, and highly correlated, which makes it difficult to handle. Moreover, underlying biological mechanisms are extremely complex and patient populations are largely diverse. As a consequence, disease progression or drug response end up being practically unique for each patient, which makes it challenging to optimize treatment or develop effective drugs. For instance, most major drugs are known to be effective in only 25 to 60 percent of patients, and more than 2 million cases of adverse drug reactions occur annually in the United States, including 100,000 deaths [218]. Such problem is aggravated in small sample-size settings, i.e., when the number of patients is smaller than the number of available observations per patient, which is often the case in clinical trials. Thus, improved statistical methods

are imperatively needed to analyze data from clinical trials, which can account for heterogeneous, noisy, and missing data. Such methods should also be able to:

(i) model population heterogeneity by sharing information across patients.

(ii) deal with correlations in high-dimensional data.

(iii) differentiate between natural response of patients and effects due to the treatment.

This chapter proposes a BNP method to discover both prognostic and predictive biomarkers, while identifying relevant subpopulations in clinical trials. Our methodology allows us to isolate drug effects, i.e., we are able to automatically differentiate between effects due to the natural response of patients and those caused by the treatment itself. At the same time, we are able to discover different subpopulations that might be characterized by different sets of biomarkers. We identify *global* biomarkers that are relevant for the whole patient population, as well as *local* biomarkers that only affect specific subpopulations. By performing differential analysis between subpopulations, we can give information on the direction of change and effect size[1] of each variable. Our method is nonparametric, which means that the number of subpopulations and sets of relevant biomarkers are automatically learned from data.

We follow a Bayesian approach to better account for missing data and uncertainty [75, 191]. Our method extends the GLFM, an approach based on the Indian buffet process (IBP) that is able to deal with mixed continuous and discrete data [72, 210]; we extend and adapt such model to deal with the case-control clinical trial scenario. The IBP has already been used in biological applications such as analysis of gene expression data [107], biological interaction networks [77], multi-platform genomics [173], genetic tumor variation [28, 112], or comorbidity analysis of psyquiatric disorders [211]. To the extent of our knowledge, this is the first time that the IBP has been applied to analyze data from clinical trials. We demonstrate the usefulness of our novel approach on a real dataset, i.e., a randomized phase II case-control study for the assessment of a cutting-edge immunotherapy treatment for liver cancer [2]. Not only our method finds already well-known relevant biomarkers, it also discovers new biomarkers that could not be found with previous methods, both at a global and subpopulation specific level [158].

4.2 General Latent Feature Model

Latent feature models are unsupervised approaches to analyze complex datasets. They model the joint probability of the data $p(\mathbf{X})$ using a set of latent features [135]. Each latent feature captures

[1] In statistics, an effect size is a quantitative measure of the strength of a phenomenon. See Section 4.3.3 for further details.

common correlation patterns among the dimensions, and the objective is to learn the most probable set of such latent features.[2] Figure 4.1 illustrates the idea underlying a latent feature model. $\mathbf{X}$ can be decomposed into the product of two matrices that should be learned from data: a feature-activation matrix (also called weight matrix) $\mathbf{Z}$ and a dictionary matrix $\mathbf{B}$. Each element x_{nd} of matrix $\mathbf{X}$ results from a linear combination of K feature elements B_{kd}, i.e., x_{nd} corresponds to the realization of a random variable following the probability distribution $f(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d})$, where $\mathbf{Z}_{n\bullet}$ is the n-th row of $\mathbf{Z}$ and $\mathbf{B}_{\bullet d}$ is the d-th column of $\mathbf{B}$. The simplest latent factor model assumes Gaussian priors over both the weight and dictionary matrices.

The GLFM, which was first introduced in [209], improves upon classical latent factor models in three aspects. First, it is a Bayesian nonparametric model where the number of latent features is also learned from data [61, 72]. In other words, the model assumes an a priori unbounded number K of latent features, usually denoted by $K \to \infty$. This is a useful property, given that the number of correlation patterns (corresponding to each latent feature) to be discovered is generally not known beforehand.

Second, the GLFM can handle heterogeneous datasets and missing observations straightforwardly. This comes handy to deal with clinical trial data, where observations for each patient are typically diverse in nature, and missing values occur frequently (e.g., not all patients might get the same tests run, others might drop out from the study at some point, etc.).

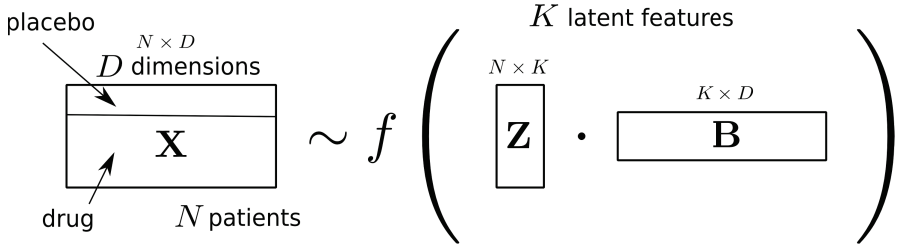


Figure 4.1: **Illustration of the matrix factorization scheme.** $\mathbf{Z}$ is the feature-activation matrix, $\mathbf{B}$ is the dictionary matrix, and f is the model likelihood which depends on the type of data of each dimension. Although we depict a single function f in the diagram, this is a slight abuse of notation, since the transformation might be different for each dimension d. We will see later that f actually corresponds to multiple link functions $T_d(\cdot\,; \boldsymbol{\phi_d})$ that will vary column-wise depending on the type of data and be applied element-wise.

The third advantage of the GLFM is that it allows for a partition of the patients in different subpopulations. The model assumes a binary feature-activation matrix $\mathbf{Z} \in \{0, 1\}^{N\times K}$, which admits an easy interpretation in which each latent feature can be either active or inactive for each patient. Patient subpopulations can then be identified by gathering all patients that have the same set of active features. Within the Bayesian framework, the GLFM resorts to the IBP as a prior for the feature-

[2] More precisely, the objective is to learn a posterior distribution for each latent feature.

activation matrix $\mathbf{Z}$ [209].

To deal with mixed continuous and discrete observations together [210], the GLFM assumes that each observation x_{nd} comes from the transformation of an auxiliary Gaussian-distributed random variable y_{nd}, resulting in the following generative model:

$$x_{nd} = T_d(y_{nd}; \boldsymbol{\phi_d}) \tag{4.1}$$

$$y_{nd}|\mathbf{Z}, \mathbf{B} \sim \mathcal{N}(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}, \sigma_y^2), \qquad \mathbf{Z} \sim \mathrm{IBP}(\alpha). \tag{4.2}$$

Gaussian priors are assumed over each element $B_{kd} \sim \mathcal{N}(0, \sigma_B^2)$. The function $T_d(\cdot\,; \boldsymbol{\phi_d})$ is the likelihood transformation for dimension d, a link function which will depend on the type of data at hand, and $\boldsymbol{\phi_d}$ are eventual parameters for the chosen transformation $T_d(\cdot\,; \boldsymbol{\phi_d})$. The inference proposed for such model follows a Markov chain Monte Carlo (MCMC) approach, namely, an accelerated Gibbs sampling (AGS). The idea of AGS is that, instead of completely collapsing $\mathbf{B}$, we instantiate it at each iteration using a subset of the data, randomly chosen and different at each iteration. Algorithm 2 shows the inference procedure for a single iteration of the GLFM.

Algorithm 2 Single iteration in the MCMC procedure for the GLFM (see [210] for further details).

Input: feature activation matrix $\mathbf{Z}$, pseudo-observations $\mathbf{Y}$, observations $\mathbf{X}$

1: update $\mathbf{Z}$ given $\mathbf{Y}$ using AGS
2: **for** $d = 1, \ldots, D$ **do**
3: sample $\mathbf{B}_d$ given $\mathbf{Z}$, $\mathbf{Y}_d$
4: sample $\mathbf{Y}_d$ given $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{B}_d$
5: **end for**

Output: activation matrix $\mathbf{Z}$, pseudo-observations $\mathbf{Y}$, and dictionary $\mathbf{B}$

4.3 Our Approach

Let us consider a clinical trial for the assessment of a certain treatment. N is the total number of patients involved in the clinical trial, and D is the number of available observations for each patient – those might include demographics, genetic data, clinical or environmental information. Let $\mathbf{X}$ be the $N \times D$ input matrix of the observations for all patients. Such matrix might be incomplete, noisy and heterogeneous (it might contain very diverse data such as continuous, positive real-valued, categorical, ordinal, or count data). Among the N patients, N_0 patients have taken a placebo (we refer to this group as the *placebo arm*), and N_1 others have received the actual drug, thus belonging to the *treatment arm*. Let $\mathbf{R} \in \{0, 1\}^{N \times 1}$ be the drug indicator vector, which takes non-zero values for patients belonging to the treatment arm.

Among the D available dimensions for each patient, we have a variable d^* (or several) that captures how well patients are doing, e.g., the elapsed time until tumor size increases.[3] In such scenario, our objective is to discover prognostic and predictive biomarkers with respect to d^*, i.e., prognostic variables that help us predict the natural evolution of patients regardless of treatment, and predictive variables to anticipate patient drug responses.

Here, we resort to an unsupervised approach[4] based on the GLFM where we model the joint probability distribution $p(\mathbf{X})$. The gist of our method for biomarker screening is to first, find an useful projection of the patients population into a latent space, and second, conduct a set of hypothesis tests based on such partition of the patients, i.e., a multiple hypothesis testing procedure to quantify the statistical significance of each biomarker. Finding an adequate basis for the latent space is crucial to reveal an useful set of subpopulations (patients with similar biomarker profiles) where drug effects are isolated from natural disease responses, thus revealing potential targets for improvement of the efficiency of the drug.

4.3.1 Modeling

In order to deal with the small sample-size scenario typical from clinical trials, we adapt the GLFM to share information between patients in the placebo and treatment arm. In particular, we allow for two types of latent features: *global* features and *treatment-specific* features. Global features are learned from patients in the placebo arm, and can be active for any patient, capturing general patterns in the patient population regardless of any treatment. In contrast, drug-specific features are learned from treated patients solely, and can only be active for patients in the treatment arm, capturing correlations linked to the effect of the drug. We call this extension the case-control Indian buffet process (C-IBP) feature model.

Let $\mathbf{Z}$ and $\mathbf{W}$ be the feature activation matrices for global and drug-specific features respectively. Note that the model will learn the necessary number of global features K and treatment-specific features K'. The complete generative model is given by

[3] The definition of variable d^* is important for the post-processing statistical analysis of the model output.

[4] Note that dimension d^* is included in matrix $\mathbf{X}$, we treat such variable in the same way as any other observation, in contrast to any supervised approach such as linear mixed models or hazard modeling. A standard supervised approach was already conducted for our application at hand at [2].

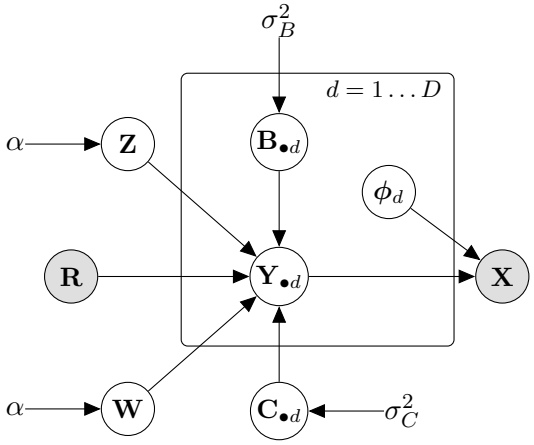


Figure 4.2: **Graphical model for the C-IBP.** Nodes represent random variables: grey ones are observed, whereas white ones are hidden. $\mathbf{X}$ is the observation matrix for all patients, and $\mathbf{R}$ is the drug indicator vector to distinguish patients in the placebo and treatment arms. $\mathbf{Z}$ and $\mathbf{W}$ are the feature activation binary matrices for global and treatment-specific features, whereas $\mathbf{B}$ and $\mathbf{C}$ are the respective dictionary matrices.

$$x_{nd} = T_d(y_{nd}; \boldsymbol{\phi_d}) \tag{4.3}$$

$$y_{nd}|\mathbf{Z}, \mathbf{W}, \mathbf{B}, \mathbf{C}, \mathbf{R} \sim \mathcal{N}(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d} + \mathbb{1}[\mathbf{R}_n = 1]\mathbf{W}_{n\bullet}\mathbf{C}_{\bullet d}, \sigma_y^2) \tag{4.4}$$

$$B_{kd} \sim \mathcal{N}(0, \sigma_B^2), \qquad C_{kd} \sim \mathcal{N}(0, \sigma_C^2) \tag{4.5}$$

$$\mathbf{Z} \sim \text{IBP}(\alpha), \qquad \mathbf{W} \sim \text{IBP}(\alpha), \tag{4.6}$$

where $\mathbf{B}$ and $\mathbf{C}$ are the dictionary matrices for global and treatment-specific features, respectively. Figure 4.2 shows a graphical representation of the C-IBP. Concerning the link functions $T_d(\cdot\,; \boldsymbol{\phi_d})$, we use the default transformations described in [210].

4.3.2 Inference

Following [210], we use an MCMC method, which has been broadly applied in other IBP-based models [72, 131, 219]. An important remark is that conditioned on $\mathbf{Y}$, the feature activation matrices $\mathbf{Z}$, $\mathbf{W}$, and the dictionary matrices $\mathbf{B}$, $\mathbf{C}$, are all independent from the observation matrix $\mathbf{X}$. In order to infer matrices $\mathbf{Z}$ and $\mathbf{W}$, we resort to a collapsed Gibbs sampler, which presents better mixing properties than the uncollapsed one. Gibbs sampling is often chosen in the context of the standard linear-Gaussian IBP because of its simplicity, but its computational cost is relatively high [72]. To make the algorithm more efficient, we resort to the AGS which is fully described in [39]. Some of the sampling steps for the feature activation matrices can then be done independently and in parallel, as in the uncollapsed case.

A simple sub-optimal inference procedure for the C-IBP can be directly derived based on the

inference for the GLFM. We first learn the global features by training the GLFM with patients belonging to the placebo arm exclusively. We thus learn a set of global features to describe general placebo patients. Next, we learn the drug-specific features for patients in the treatment arm. This is performed by training the GLFM model with the whole patient population and imposing the following constraints:

(i) global features (corresponding to matrix $\mathbf{B}$) are kept fixed to the values learned previously.

(ii) feature activations for patients in the placebo arm are initialized to their previous value.

(iii) drug-specific features are forced to be inactive for all patients in the placebo arm, i.e., these features are learned solely based on patients belonging to the treatment arm. These allow us to completely isolate the effect of the drug.

More precisely, let $\mathbf{Z}^0 \in \{0,1\}^{N_0 \times K}$ and $\mathbf{Z}^1 \in \{0,1\}^{N_1 \times K}$ be the global feature activation matrices for patients in the placebo and treatment arm respectively, where K is the number of inferred global features, N_0 is the number of patients in the placebo arm, and N_1 is the number of patients belonging to the treatment arm. Also, let us define $\mathbf{Y}^0$ and $\mathbf{Y}^1$ as the auxiliary variables for placebo and drug patients respectively. Patients in the placebo arm are only affected by matrices $\mathbf{Z}^0$, $\mathbf{Y}^0$, and $\mathbf{B}$. For these latent variables, we can directly apply the inference procedure described in Algorithm 2. Once $\mathbf{Z}^0$, $\mathbf{Y}^0$, and $\mathbf{B}$ have been sampled, global feature activations for drug patients $\mathbf{Z}^1$ can be sampled as

$$p(Z^1_{nk} = 1 | \mathbf{Z}_{-nk}, \mathbf{B}) \propto \frac{m_k}{N} \prod_{d=1}^{D} \mathcal{N}(y_{nd} | \sum_j Z^1_{nj} B_{jd}, \sigma_y^2), \tag{4.7}$$

where m_k is the number of patients for which feature k is active. Note that at this stage, we do not allow for the creation of new global features: we only want to know the assignment of global features for the drug population. Given all feature activation matrices $\mathbf{Z}$ and $\mathbf{W}$, $\mathbf{C}$ and $\mathbf{Y}^1$ can be sampled in the same way as $\mathbf{B}$ and $\mathbf{Y}^0$ before. The whole inference procedure is summarized in Algorithm 3.

4.3.3 Statistical Methodology

Once the model has been trained (samples from an approximate posterior distribution can be drawn), we proceed with a classical frequentist approach[5] to identify statistically significant prognostic and predictive biomarkers. The whole procedure is summarized in Algorithm 4. First, we take M posterior samples from the posterior distribution of $\mathbf{Z}$. For each sample, patients that have the same

[5] Although Bayesian approaches to quantify statistical significance exist, such as posterior predictive checks or Bayesian factors, classical statistics predominate in the bio-medical field.

Algorithm 3 Sub-optimal inference procedure for the C-IBP model.

Input: observation matrix $\mathbf{X}$

Initialize: feature activation matrices $\mathbf{Z}$ and $\mathbf{W}$, and pseudo-observations matrix $\mathbf{Y}$
for each iteration **do**
  sample $\mathbf{Z}^0$, $\mathbf{Y}^0$, and $\mathbf{B}$ given $\mathbf{X}$, according to Algorithm 2
  for $d = 1, \ldots, D$ **do**
    sample $\mathbf{Z}^1$ given $\mathbf{Y}^1$ and $\mathbf{B}$ according to (4.7)
  end for
  sample $\mathbf{W}$ given $\mathbf{Z}^1$ and $\mathbf{Y}^1$ using AGS
  for $d = 1, \ldots, D$ **do**
    sample $\mathbf{C}_d$ given $\mathbf{Z}$, $\mathbf{W}$, and $\mathbf{Y}_{\bullet d}$
    sample $\mathbf{Y}^1_d$ given $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{W}$, $\mathbf{B}_{\bullet d}$ and $\mathbf{C}_{\bullet d}$
    sample $\boldsymbol{\phi_d}$ if needed
  end for
end for

Output: feature activation matrices $\mathbf{Z}$ and $\mathbf{W}$

activation pattern of features can be grouped together in the same subpopulation. For instance, subpopulation (1001) refers to all patients having the first and forth feature active. Let P refer to the total number of inferred subpopulations across the M posterior samples. By considering multiple posterior samples, we obtain slightly different partitions of patients in subpopulations. This can be seen as performing *soft-clustering* of patients, i.e., patients that are in-between subgroups might be assigned to different subpopulations in different posterior samples. Thus, the method is more robust against model inaccuracies at clustering patients. This is an important benefit of Bayesian modeling in general.

Next, in order to also make our method robust against outliers (patients with extreme biomarker values), we perform bootstrapping L times, for each subpopulation and posterior sample. Bootstrapping relies on random sampling with replacement. It is a technique used for computing robust estimators against outliers by sampling from an approximating distribution, which is particularly useful for hypothesis testing when the model assumptions are in doubt or unknown [216]. The standard bootstrapping approach relies on the construction of an estimator for hypothesis testing based on a number L of resamples with replacement of the observed dataset (and of equal size to the observed dataset), i.e., sampling with replacement from the empirical distribution of the observed data.

Given M posterior samples and L bootstrapping instances for each sample, we end up with ML different subpopulation instances. Measures of effect size (quantitative measure of the difference between two subpopulations) and statistical significance can be computed for each instance and then averaged across them, so that partition inaccuracies and outlier effects are mitigated. In the described

algorithm, we suggest to compare each possible pair of subpopulations,[6] but we might want to focus only on the biggest communities or specific subpopulations of interest to reduce computational cost. Let Q be the total number of considered comparisons between subpopulations. In our particular case, $Q = P \cdot (P-1)/2$ as we consider each pairwise comparison among the P subpopulations. Let $i(q)$ and $j(q)$ refer to the set of subpopulation indexes corresponding to comparison q, e.g., $i(q) = 4$ and $j(q) = \{1, 2, 3\}$ corresponds to the comparison of subpopulation 4 against subpopulations 1, 2, and 3 aggregated. In the following, we will describe how to compute the $Q \times D$ effect size and statistical significance matrices.

Algorithm 4 Statistical approach for biomarker discovery (post-processing procedure).

Input: M posterior samples from $\mathbf{Z}$ and $\mathbf{W}$, list of P subpopulations, and Q comparisons

1: **for** $m = 1, \ldots, M$ **do**
2: bootstrap for each subpopulation L times
3: **end for**
4: **for** $q = 1, \ldots, Q$ **do**
5: choose subpopulations $G^{i(q)}$ and $G^{j(q)}$
6: compute effect size according to Eq. 4.8, 4.9, and 4.10.
7: compute statistical significance (p-value) according to the Mann-Whitney test for continuous variables and Fisher test for discrete variables, adjusting for multiple hypothesis testing [17]
8: **end for**

Output: effect size matrix $\mathbf{\Delta}$ and significance matrix $\mathbf{\Upsilon}$, both of dimensions $Q \times D$

Effect size. For each comparison q and dimension d, we compute the effect size Δ_{qd} as:

$$\Delta_{qd} \quad = \quad \mathbb{E}_{m,l}\left[\delta_{qd}(m,l)\right], \tag{4.8}$$

where $\boldsymbol{\delta}_{qd}$ is an $M \times L$ matrix of relative effect sizes for each posterior sample m and bootstrap iteration l. The expectation is done across all posterior samples and bootstrapping iterations, which are equally probable. In the case of continuous variables, we define

$$\delta_{qd}(m,l) \quad = \quad \log_2\left(\frac{\mu_d\left(\mathcal{G}_{ml}^{i(q)}\right)}{\mu_d\left(\mathcal{G}_{ml}^{j(q)}\right)}\right), \tag{4.9}$$

where $\mathcal{G}_{ml}^{i(q)}$ and $\mathcal{G}_{ml}^{j(q)}$ refer to subpopulations $i(q)$ and $j(q)$ in the posterior sample m and bootstrap iteration l, and $\mu_d(\mathcal{G})$ is the mean value of variable d within a given subpopulation $\mathcal{G}$. Taking the logarithm facilitates interpretation, such that an increase or decrease ratio has the same scale: for

[6] Other comparison schemes are possible, such as a leave-one-out strategy consisting in the comparison of each individual subpopulation against the rest. Note that as the number of comparisons increase, the correction for multiple hypothesis testing shall be stronger.

instance, $\delta_{qd}(m,l) = 0$ means that variable d has exactly the same averaged value in both subpopulations, $\delta_{qd}(m,l) = +1$ means that variable d is twice higher in subpopulation i, and $\delta_{qd}(m,l) = -2$ means that variable d is four times smaller in subpopulation $\mathcal{G}_{ml}^{i(q)}$ with respect to subpopulation $\mathcal{G}_{ml}^{j(q)}$. In the case of a discrete variable d, we check for mean differences, i.e.,

$$\delta_{qd}(m,l) = \mu_d(\mathcal{G}_{ml}^{i(q)}) - \mu_d(\mathcal{G}_{ml}^{j(q)}). \tag{4.10}$$

Note that we define different measures for continuous and discrete variables as the dynamic range of continuous variables is generally much higher, making the logarithmic scale more appropriate.

Statistical significance. To measure how significant an effect size $\delta_{qd}(m,l)$ is, for each posterior sample m and bootstrap instance l, we compute a statistical significance value $v_{qd}(m,l)$ as the p-value resulting from a certain two-sample test. In general, selecting the most appropriate statistical test in hypothesis testing is a challenging task [94, 127]. Here, we opt for one statistical test for all continuous variables and another one for discrete variables for simplicity, although more sophisticated strategies could certainly be investigated. We use the Mann-Whitney test for continuous variables and the Fisher test for discrete variables. The Mann-Whitney test is a general nonparametric statistical test to check whether the distribution of both populations are equal without requiring any normality assumption. The Fisher test is a standard test for categorical variables [216]. We define the $Q \times D$ matrix of statistical significance $\boldsymbol{\Upsilon}$, for each comparison q and biomarker d as the median p-value across the M samples and L bootstrapping instances:

$$\Upsilon_{qd} = \text{median}_{m,l}\left[v_{qd}(m,l)\right], \tag{4.11}$$

where $\boldsymbol{v}_{qd}$ denote the $M \times L$ matrix of statistical significance values $v_{qd}(m,l)$ for each posterior sample m and bootstrapping instance l. Finally, we follow the Benjamini Hochberg procedure for multiple hypothesis testing to adjust the statistical significance threshold α_s such that a certain false discovery rate (FDR) is guaranteed [17]. A biomarker d is said to be statistically significant for comparison q if its significance value Υ_{qd} (the median p-value across posterior samples and bootstrapping instances) is smaller than the adjusted threshold, i.e., $\Upsilon_{qd} < \alpha_s$.

4.4 Results

To evaluate the performance of our method, we consider a randomized phase II trial regarding an immunotherapy treatment against liver cancer [2]. This clinical trial studies the effect of Codrituzumab,

a manufactured antibody treatment against a liver cancer protein called Glypican-3 (GPC3) that is expressed in hepatocellular carcinoma (HCC). GPC3 is a member of the glypican family, a group of heparan sulfate proteoglycans linked to the cell surface and which plays an important role in cell growth, differentiation, and migration [116, 52]. GPC3 is highly expressed in HCC and has become a useful diagnostic marker for HCC by immunohistochemical (IHC) studies, since the adjacent non-tumoral tissue does not express GPC3 [221]. GPC3 may promote HCC growth by stimulating the canonical Wnt pathway, and/or interacting with the IGFII-IGF1R pathway, or it may play a role in FGF signaling [26]. Therefore, GPC3 may represent a specific tumor marker and a potential target for therapy in HCC [138].

4.4.1 Antibody Treatment for Hepatocellular Carcinoma

Codrituzumab is a recombinant, humanized monoclonal antibody that binds to human GPC3 with high affinity [98, 197]. Codrituzumab interacts with CD16/FcγRIIIa and triggers antibody-dependent cytotoxicity (ADCC), as shown by [137]. Non-clinical characterization of Codrituzumab demonstrates that it elicits ADCC against GPC3-positive human hepatoma cell lines, using human peripheral blood mononuclear cells as effector cells [198]. Phase I studies in US and Japan showed that Codrituzumab was well tolerated up to 20 mg/kg/wk without dose limiting toxicity [227, 93].

Here, the considered phase II clinical trial aims at comparing Codrituzumab versus placebo in advanced HCC patients who had failed prior systemic therapy (whole body treatment). The database consists of 180 patients, 60 patients in the placebo arm and 120 patients in the treatment arm. For each patient n, we consider 71 observations which include demographic information, characteristics of the tumor and diverse clinical measurements. We used progression free survival (PFS) as a clinical endpoint in our analysis. This variable measures the elapsed time in months until the tumor grows in size, which is correlated to the survival of HCC patients.

According to a first analysis of this database published in [2], Codrituzumab was not found to be effective against liver cancer, although it was suggested that a higher dose of Codrituzumab or selecting patients with high level of GPC3 could improve the outcome. This directly relates to the patient heterogeneity problem explained in Section 4.1. By directly addressing this issue, our approach gives us additional information that could not be extracted in that first study.

4.4.2 Identified Subpopulations

As shown in Table 4.1, the C-IBP model identified twelve subpopulations from the set of 180 patients, and three latent features (F1, F2, and F3) which capture correlation patterns of biomarker values.

Sub-population	Drug Identifier	F1	F2	F3	Average num. of patients	Mean PFS (months)	Median PFS (months)
1.	0	0	0	0	33.37	3.06	1.65
2.	0	0	1	0	4.07	2.29	2.24
3.	0	1	0	0	17.84	2.72	1.81
4.	0	1	1	0	4.72	7.05	7.18
5.	1	0	0	0	51.52	3.22	2.55
6.	1	0	0	1	16.77	4.17	3.65
7.	1	0	1	0	8.38	1.74	1.33
8.	1	0	1	1	2.07	2.69	2.65
9.	1	1	0	0	29.88	3.36	2.03
10.	1	1	0	1	4.90	4.44	4.34
11.	1	1	1	0	4.53	6.31	5.31
12.	1	1	1	1	1.94	10.04	10.01
Total	120.0	63.82	25.72	25.69	180	3.44	2.04

Table 4.1: **List of subpopulations identified by the C-IBP.** Each subpopulation (row) is represented by a feature activation pattern, e.g., (1100). Last row respectively corresponds to the average number of patients having each feature active, as well as the total size of the database, global mean and median PFS values.

These features can be either active ($= 1$) or inactive ($= 0$) for each patient individually. A feature is active when the corresponding pattern contributes to the total biomarker values of that patient. A subpopulation is defined as a group of patients having similar biomarker values, encoded by the same set of active features. Among the three inferred features, F1 and F2 are global, and F3 is a drug-specific feature. The later captures variations exclusively due to the drug, as it can only be active for patients in the placebo arm, corresponding to subpopulations 1 to 4. The last three columns in the table tell us how many patients are present on average in each subpopulation,[7] as well as the mean and median values of the PFS.

In the placebo arm, the C-IBP model inferred four subpopulations from data (groups 1 to 4 in Table 4.1). Subpopulation 4 has a significantly higher PFS, indicating better prognosis. In the treatment arm (groups 5 to 12 in Table 4.1), subpopulations with feature F3 active (6, 8, 10 and 12) have higher PFS values. The effect of F3 on prolonged survival was found across all subpopulations, as seen by pairwise comparisons between subpopulations in the treatment arm with F3 active vs. the ones with F3 inactive (e.g., 5 vs 6, 7 vs 8, etc).

Table 4.3 shows a boxplot illustration for PFS values within each of the twelve inferred subpopulations. Within the placebo arm, subpopulation 4 is the only sub-group that has much longer PFS values or survival, whenever both F1 and F2 are active. Concerning the treatment arm, the last three

[7]Note that the size of each subpopulation is not an integer number of patients because we are averaging over different posterior samples and bootstrapping iterations at each time, as discussed in Section 4.3.3

communities (subpopulations 10, 11, and 12) have PFS values that are more than twice higher with respect to the global average. Yet, it is not clear a priori whether such communities have longer PFS because of the drug, or because they had a good natural predisposition from the beginning. Latent features can help us answer this question. In particular, by looking at the global features F1 and F2, we can see that communities 11 and 12 share the same activation pattern than subpopulation 4; this means that they share the same biomarker values/good conditions for a positive natural response, independently of the drug. In contrast, community 10 has the same activation pattern than community 3 for the global features, which indicates that there is no a priori reason to expect a positive natural response from these patients: the survival increase in that case would be exclusively due to the drug, and captured by feature F3. In general, note that whenever feature F3 is active, there is a systematic increase in PFS values for each community with respect to their equivalent subpopulation with F3 equal to zero. Thus, we can say that F3 captures a positive effect of the drug on patients, which is stronger whenever feature F2 is also active. Figure 4.3 shows the survival or PFS values as box plots for each inferred subpopulation.

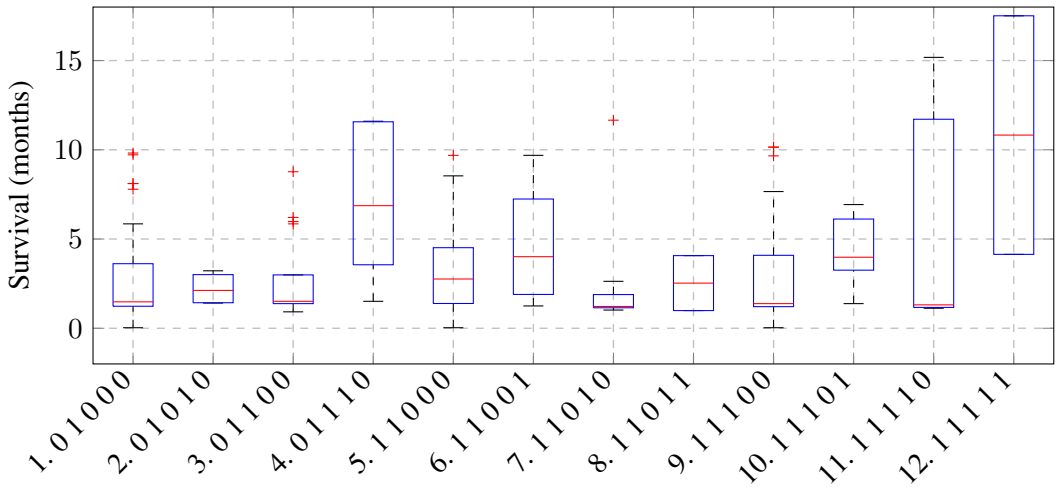


Figure 4.3: **Survival boxplots for each subpopulation.** We represent the distribution of PFS (in months) for each subpopulation. The patterns in the x-axis correspond to those listed in Table 4.1.

4.4.3 Discovered Biomarkers

Next, we analyze the effect of each latent feature activation on the biomarker values. Figure 4.4 depict the relative mean effect size of each biomarkers in relation to each of the three latent features (F1-F3). The black vertical lines correspond to two standard deviations of the relative effect size matrix $\boldsymbol{\delta}_{qd}$. Biomarkers which are statistically significant at significance level $\alpha_s = 0.001$ are marked with a red

circle. Table 4.2 corresponds to the same information in numerical values, as well as p-values for each biomarker. More precisely, these figures and tables result from the comparison of aggregated subpopulations in which one particular feature is active against all the other subpopulations (where that particular feature is inactive). For instance, Figure 4.4a results from the statistical procedure described in 4.3.3 applied to subpopulations (3,4,9,10,11,12) against (1,2,5,6,7,8). The sign (+ or -) is important for the biological interpretation of the data, and indicates the direction of the biomarker effect on the case group versus the control group, such that positive effect means that the biomarker has a higher value in the case group versus the control group.

Variable description	Effect size	Mw-test
AFP	3.65	**1.31e-07**
H score Mem	1.30	**1.53e-21**
H score Cyt	0.69	**2.08e-16**
sGPC3 11/96	0.47	**1.34e-04**
sGPC3 30/57	0.40	**1.53e-05**
sGPC3 30/607	0.34	**6.21e-10**
sGPC3 114/165	0.32	**2.48e-10**
CD56-CD16+	0.34	**1.42e-03**
CD56dimCD16-	0.22	**7.69e-04**
NK	0.42	**4.40e-03**
CD56dimCD16 bright	0.42	**6.23e-03**
CD16	0.10	**3.88e-03**
NKP46	0.09	**3.08e-03**
CD8	0.07	**1.41e-02**
CD3	0.05	**5.38e-03**
CD45	0.05	**6.88e-03**

Variable description	Effect size	F-test
IHC 3+	0.84	**1.96e-31**
IHC 1+	-0.42	**9.97e-11**
IHC 2+	-0.23	**3.67e-06**
IHC 0	-0.19	**4.42e-05**
FCGRIIIA-158-C/A	-0.25	**7.16e-04**
FCGRIIIA-158-A/A	0.20	**9.21e-03**

(a) Feature F1

Variable description	Effect size	Mw-test
CD3/CD16 necrotic / stroma	3.94	**1.29e-17**
P Necrotic / stroma	3.91	**9.39e-18**
CD3/CD16 tumor viable	-1.27	**1.29e-17**
P Tumor Viable	-1.24	**9.39e-18**
sGPC3 30/607	-0.25	**4.02e-04**
AAT	-0.58	**1.54e-03**
sGPC3 114/165	-0.21	**2.43e-03**

Variable description	Effect size	F-test
-	-	-

(b) Feature F2

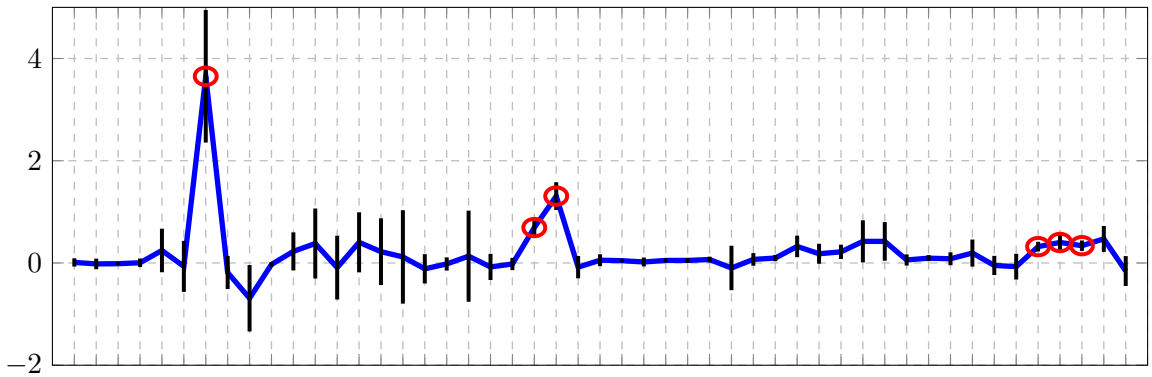
Variable description	Effect size	Mw-test
CD56dimCD16 bright	1.27	**2.18e-09**
NK	1.24	**2.13e-10**
CD56-CD16+	0.61	**2.19e-06**
CD56dimCD16-	0.42	**1.04e-05**
CD8 NK	0.37	**7.13e-08**
DN	0.27	**3.73e-05**
CD16	0.24	**6.06e-10**
NKP46	0.22	**1.14e-09**
CD8	0.20	**3.44e-09**
CD45	0.14	**5.09e-11**
CD3	0.13	**1.49e-08**
CD4	0.12	**5.06e-06**
H score Mem	-0.89	**1.73e-03**
DP	0.54	**3.55e-04**
B	0.19	**8.55e-04**
CD56	0.18	**3.05e-03**
AFP	-3.34	**1.86e-02**
P Stroma	-1.26	**6.70e-03**
CD4/CD8	-0.75	**9.05e-03**
Ctrough	0.59	**4.24e-03**
P Tumor	0.35	**1.10e-02**
sGPC3 30/607	-0.19	**1.59e-02**

Variable description	Effect size	F-test
-	-	-

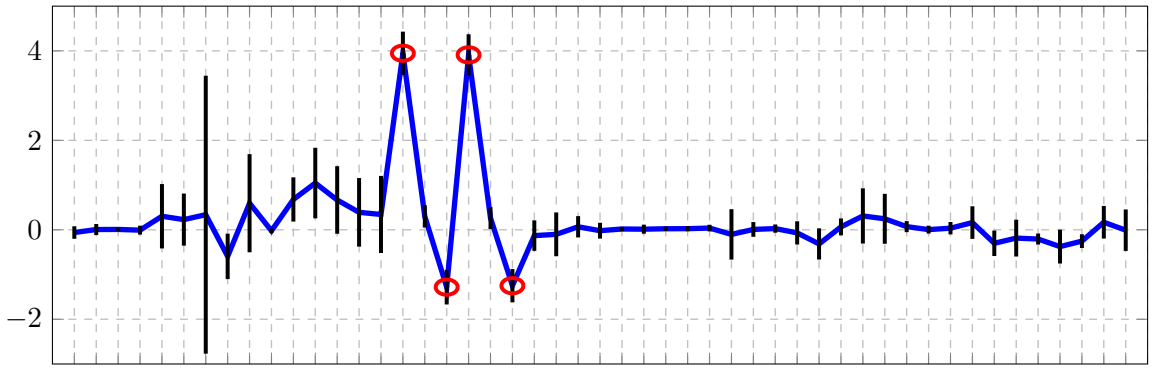
(c) Feature F3

Table 4.2: **Significant biomarkers associated to the activation of each inferred latent feature.** F1 and F2 are global features, whereas F3 is a treatment-specific feature. Colors code for the different statistical significance levels: $\alpha_s = 0.001$ (black), $\alpha_s = 0.005$ (blue), and $\alpha_s = 0.01$ (green). Continuous variables on the top, discrete variables on the bottom.

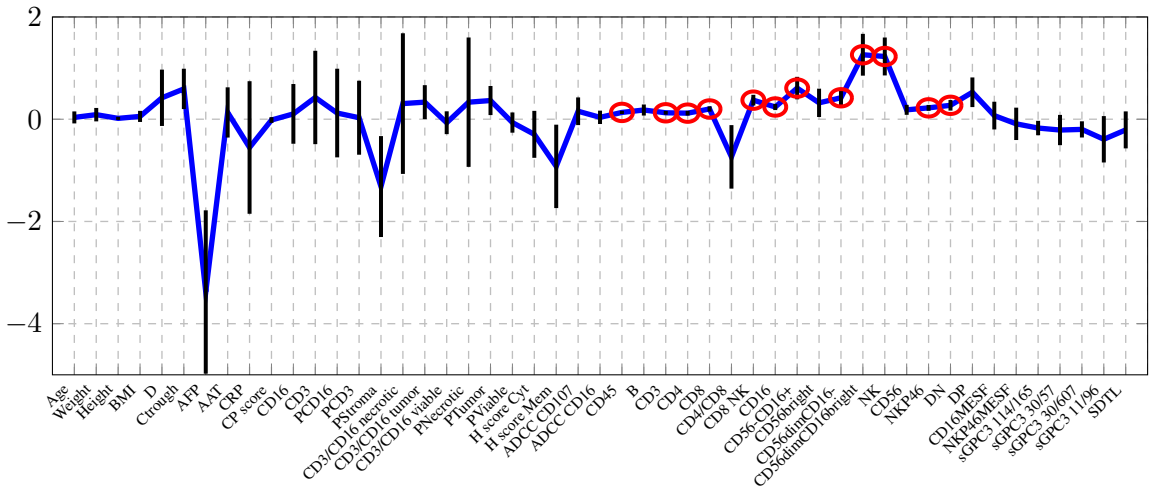
Feature F1 in Figure 4.4a is associated with high levels of alpha-fetoprotein and GPC3 expression both in tumor (GPC3 cytoplasmic and membrane IHC staining), and as soluble protein sGPC3 (as detected by four serum assays). Feature F2 is associated with higher levels of inflammatory T and NK cells (positively stained for CD3/CD16) in tumor necrotic tissue and adjacent peri-tumoral stroma,



(a) F1



(b) F2



(c) F3

Figure 4.4: **Relative effect size of biomarkers associated to each latent feature inferred by the C-IBP model.** Significant biomarkers according to the Mann-Whitney test are marked with red circles ($\alpha_s = 0.001$). F1 identifies two types of patients with similar prognosis but different characteristics, F2 and F3 are associated with higher Progression Free Survival: F2 capture prognostic biomarkers while F3 capture predictive biomarkers.

and low levels of CD3/CD16 in viable tumor cells, as shown in Figure 4.4b. Feature F3 is associated with higher levels of T (CD3, CD45) and NK (CD16, NKp46) cell markers, and low levels of alpha-fetoprotein (see Figure 4.4c).

4.4.4 Discussion

The identification of biomarkers in complex datasets affected by multiple confounding factors can be challenging. In drug development, clinical trials are powered to deal with situations of constant drug exposure, and traditional statistical methods have difficulties extracting signals from data confounded by factors like differential drug exposure. Therefore, sensitive analytical methods are needed to deconvolute real signals from biological and technical noise. We applied the C-IBP to a phase II clinical study of Codrituzumab in HCC which had failed to meet the primary endpoint due to insufficient drug exposure in the treatment arm. Traditional statistical approaches did not render useful insights into potential biomarkers of response; in contrast, the IBP-based analysis identified several biomarkers that stratified patient subgroups with statistical significance.

The C-IBP model identified two kinds of features: *global* features, which were active for patients regardless of treatment and indicated prognostic markers, and *drug-specific* features, which were active for patients only in the treatment arm. Global prognostic features included known prognostic markers in HCC, like alpha-fetoprotein and GPC3 expression in the tumor. In addition, global features included levels of inflammatory T and NK cells in tumor necrotic tissue and adjacent peri-tumoral stroma. Patients with higher levels of CD3/CD16 staining in peri-tumoral tissue and lower levels in viable tumor cells had better prognosis, which is consistent with the role of inflammatory cells in anti-tumor response [57]. Drug-specific features included different NK cell subtypes. This is consistent with the mode of action of Codrituzumab, which requires engagement of the CD16/FcγRIIIa receptor in NK cells to recruit NK cells to the tumor and subsequent tumor lysis [137]. The C-IBP method allows for flexible interrogation of the data, for example, the effect of F3 on prolonged survival was found across all subpopulations, as seen by pairwise comparisons between subpopulations in the treatment arm with F3 active vs. the ones with F3 inactive (e.g., 5 vs 6, 7 vs 8, etc. . . in Table4.1).

4.5 Summary

In this chapter, we have presented a BNP method for subpopulation characterisation and biomarker discovery in clinical trials. The BNP is very flexible, as it automatically infers the required number

of latent features that better explain the observations. We were able to identify both prognostic and predictive variables, and quantify the direction of action, effect size and statistical significance for each biomarker. Our model handles noise and missing information naturally, as well as heterogeneity in the types of data (continuous versus discrete).

Through the use of global and drug-specific features, the C-IBP model clearly separates between drug effects and natural prognostic factors. Also, our methodology of combining a Bayesian model with classical statistical tools is robust in several aspects: first, instead of doing hard-clustering, we compute a soft-partition by using the average of all feature activations in the posterior. Second, we combine bootstrapping techniques with posterior samples of the model, and statistical significance is assessed using classical well-known statistical tests. In summary, the C-IBP applied to a complex phase II clinical study in HCC confounded by several factors was able to identify prognostic and predictive biomarkers of response to Codrituzumab. Given the large heterogeneity in response to cancer therapeutics, novel methods of identifying mechanisms and biology of variable drug response and ultimately treatment individualization will be indispensable [148].

4.A Appendix: General Latent Feature Modeling Toolbox

GLFM is a software library to perform latent feature modeling in heterogeneous datasets, where the attributes describing each object can be either discrete, continuous or mixed variables. Up to our knowledge, this library provides the first available software for latent feature modeling in heterogeneous data, and includes functions for the two main applications of the GLFM, i.e., missing data estimation (a.k.a. table completion) and data exploratory analysis [210]. The core algorithm is developed in C++ and includes user interfaces in both Python and Matlab.[8]

4.A.1 Implementation

The main function of the package, `hidden = GLFM.infer(data)`,[9] runs the inference algorithm given the input structure `data` and returns the learned latent variables in the output structure `hidden`. This function receives as input an observation matrix $\mathbf{X}^d$ and a vector indicating the type of data for each dimension (optionally, model hyperparameters and simulation settings can be customized by the user). The latent variables are learned by making use of the mapping transformations listed in Table 4.3 to account for both continuous and discrete data types. Here, the parameters μ and w are used to shift and scale the raw input data, and are respectively set to the empirical mean and the standard deviation for real-valued attributes, and to the minimum value and the standard deviation for positive real-valued and count attributes. This guarantees that the prior distributions on the latent variables are equally good for all the attributes in the dataset, regardless their support. The output structure `hidden` contains the latent feature vectors $\mathbf{Z}_{n\bullet}$ for $n = 1, \ldots, N$, the weighting vectors $\mathbf{B}_{\bullet d}$, as well as auxiliary variables for both the likelihood and link function. Our implementation of the GLFM makes use of the GNU Scientific Library (GSL),[10] to efficiently perform a large variety of mathematical routines such as random number generation, and matrix or vector operations.

4.A.2 Usage

Data preprocessing and initialization. A convenient property of the GLFM package is that it can be used blindly on raw data without requiring any preprocessing step on the dataset, nor special tunning of hyperparameters. The only requirement for the user to use the package is to format the data as a numerical matrix of size $N \times D$ and indicate in an additional vector the type of data for each of the D attributes. As mentioned above, the parameters of the transformations in Table 4.3 are

[8] GLFM code is publicly available in `https://github.com/ivaleraM/GLFM`

[9] This call corresponds to a python call. The equivalent call in Matlab is `hidden = GLFM_infer(data)`.

[10] `https://www.gnu.org/software/gsl/`

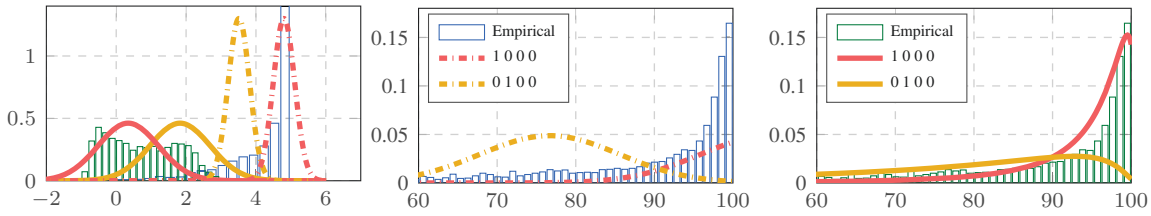
Type of Variable	Domain	Transformation $x = f_d(y)$
Real-valued	$x \in \Re$	$x = w(y+u)+\mu$
Positive real-valued	$x \in \Re^+$	$x = \log(\exp(w(y+u)+\mu)+1)$
Categorical	$x \in \{1,2,\ldots,R\}$ (unordered set)	$x = \arg\max_{r\in\{1,\ldots,R\}} y_r$
Ordinal	$x \in \{1,2,\ldots,R\}$ (ordered set)	$x = \begin{cases} 1 & \text{if } \; y \le \theta_1 \\ 2 & \text{if } \; \theta_1 < y \le \theta_2 \\ \quad\vdots & \\ R & \text{if } \; \theta_{R-1} < y \end{cases}$
Count	$x \in \{1,2,3,\ldots\}$	$x = \lfloor \log(\exp(w(y+u)+\mu)+1) \rfloor$

Table 4.3: **Mapping functions implemented in the toolbox.**

internally fixed to ensure that the pseudo-observation for all attributes fall in a similar interval of the real line, so that the prior distribution on the latent variables is equally good for all attributes.

However, we incorporate an additional functionality that allows the user to specify external pre-processing functions to further improve the performance of the algorithm. For instance, in cases in which the distribution of an attribute presents a clearly non-Gaussian behavior, e.g., it is concentrated around a single value or heavy-tailed, it might be suitable to pre-process this variable by applying a logarithmic transformation, as shown Figure 4.5.

Missing data estimation. GLFM can be used for estimation and imputation of missing data in heterogeneous datasets, where the missing values can be encoded with any (numerical) value that the user specifies. The Bayesian nature of the GLFM allows to efficiently infer the latent feature representation of the data using the available information (i.e., the non-missing values), and using it to compute the posterior distribution of each missing value in the data. Note that given the posterior distribution of each missing value, one might opt for different approaches to impute missing values, e.g., one might opt for imputing a sample of the posterior distribution or simply the maximum a posteriori (MAP) value. The GLFM package provides the function `[Xmap, hidden]=GLFM.complete(data)` which infers the latent feature representation, given the (incomplete) observation matrix, and returns a complete matrix where the missing values have been imputed to their MAP value. This function therefore runs the inference function `GLFM.infer()`, as well as the function `GLFM.computeMAP()`, which computes the MAP of a single missing element x_n^d given $\mathbf{z}^n$ and $\mathbf{B}^d$.



(a) Pseudo-observations $\mathbf{y}^d$ (b) $\mathbf{X}^d$ without pre-processing (c) $\mathbf{X}^d$ with log pre-processing

Figure 4.5: **Illustration of optional data pre-processing.** Panel (a) shows the histogram a heavy-tailed attribute and panel (b) the attribute after a logarithmic transformation, as well the distribution of the inferred latent feature patterns. Here we observe that the distribution of the attribute is better captured by the latent model when a pre-processing step is performed to correct/minimize the non-Gaussian behavior of the attribute.

Data exploration analysis. GLFM can also be used as a tool for data exploratory analysis, since it is able to find the latent structure in the data and capture the statistical dependencies among the objects and their attributes in the data. GLFM provides (weighted) binary latent features, easing their interpretation and making it possible to cluster the objects according to their activation patterns of latent features. Moreover, it also allows to activate a latent feature that is active for all the objects (a.k.a. bias term), which might be useful to capture the mode of the distribution of each attribute in the dataset. In order to ease data exploration, GLFM provides the function `GLFM.plotPatterns()`, which plots the posterior distribution for of each attribute under the given latent feature patterns, and therefore, allows us to find patterns and dependencies across both objects and attributes. This function, in turn, makes use of the function `GLFM.computePDF()`, which evaluates the posterior distribution of an attribute under a given latent feature vector.

4.A.3 Showcase Example: Voters Profile in Presidential Election

In this subsection, we illustrate the usefulness of the GLFM for data exploration in the context of politics, to identify meaningful demographic profiles, together with their geographic location, and voting tendencies in the United States. This showcase example shows how to include the specific domain knowledge into the proposed GLFM to ease the data exploration process. We apply the GLFM library to understand the correlations between demographic profiles and political vote tendencies. In particular, we focus on the United States presidential election of 1992, in which three major candidates ran for the race: the incumbent republican president George H. W. Bush, the democratic Arkansas governor Bill Clinton, and the independent Texas businessman Ross Perot. In 1992, the public's concern about the federal budget deficit and fears of professional politicians allowed the independent candidacy of billionaire Texan Ross Perot to appear on the scene dramatically [9], to the

point of even leading against the major party candidates in the polls during the electoral race.[11] The race ended up with the victory of Bill Clinton by a wide margin in the Electoral College, receiving 43% of the popular vote against Bush's 37.5% and Perot's 18.9% [109]. These results are noted for being the highest vote share of a third-party candidate since 1912, even if Perot did not obtain any electoral votes [109].

Our primary objective in this section is to find and analyze the different types of voters' profiles, as well as which candidate each profile tends to favor. To this aim, we used the publicly available *counties dataset* gathering diverse information about voting results, demographics and sociological factors per counties.[12] This dataset contains information for 3141 counties. Table 4.4 lists the per-county attributes that we used as input for our model.

Attribute description	Type of data
State in which the county is located	Categorical with 51 categories
Population density in 1992 per squared miles	Positive real data
% of white population in 1990	Positive real data
% of people with age above 65 in 1990	Positive real data
% of people above 25 years old with bachelor's degree or higher	Positive real data
Median family income in 1989 (in dollars)	Count data
% of farm population in 1990	Positive real data
% of votes cast for democratic president	Positive real data
% of votes cast for republican president	Positive real data
% of votes cast for Ross Perot	Positive real data

Table 4.4: **List of considered attributes regarding the United States presidential election of 1992.** Attributes 1 to 7 include demographic information and sociological factors, while the last three attributes summarize the percentage voting outcome in each county.

Active Feature	F1	F2	F3	F4	F5
Empirical Prob.	0.4874	0.2703	0.2700	0.0411	0.0372

Table 4.5: **Empirical feature activation probabilities for the counties dataset.** We show the empirical probability of having at least one latent feature. These are directly computed from the inferred IBP matrix $\mathbf{Z}$.

Patterns	(000)	(100)	(101)	(010)	(110)
Empirical Prob.	0.2636	0.2407	0.1063	0.1060	0.0748

Table 4.6: **Empirical probability of pattern activation for the top-five most popular patterns.** These are computed directly from the inferred IBP matrix $\mathbf{Z}$. Features F4 and F5 are always switched off, and are thus omitted from the labels.

[11] New York Times: `http://www.nytimes.com/1992/06/11/us/the-1992-campaign-on-the-trail-poll-gives-perot-a-clear-lead.html`

[12] Database available at: `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/counties.html`

Experimental setup. We run our inference algorithm with $\alpha = 5$, $\sigma_B^2 = 1$, $\sigma_\theta^2 = 1$ and the mapping transformation from the real numbers to the real positive numbers: $f(x) = \log(w \cdot (x - \mu) + 1)$, with $\mu = \mathrm{E}[\mathbf{x}^d]$ and $w = 2/\max(\mathbf{x}^d)$. In this experiment, we activate the bias term and sample the variance of the pseudo-observations for each dimension/attribute. A challenging aspect of this database is that the distributions of some of its attributes are heavy-tailed, leading to a large number of latent features as output of the GLFM, whose purpose is to capture the tails of the distributions. This is not an issue for estimation and imputation of missing data, but it renders data exploration more tedious. To solve this limitation, we here perform an additional data preprocessing step by applying a logarithmic transformation to heavy-tailed attributes. In more detail, we apply the function $g_1(x) = \log(x + 1)$ for population density, median family income, and percentage of farm population. For the percentage of white population, we used the function $g_2(x) = \log((100 - x) + 1)$ since the distribution has most of its density close to 100%.

Results. Running the GLFM on this data results in 5 latent features, whose empirical activation probabilities are shown in Table 4.5. Here, we observe that while the first three features are active for at least 27% of the counties, the last two features are active only for around 4% of the counties. Moreover, we find that the different combinations of the three first latent features represent more than 92% of the counties in USA. In the following, we will thus focus on the analysis of the three first features and, in particular, of the top-five most popular feature patterns. We show in shown in Table 4.6 the empirical probabilities of these five patterns, which represent around 80% of the U.S. counties. Figure 4.6 shows the distribution of vote percentage per candidate associated to each of these top-five patterns, while Figure 4.8 shows the corresponding geographic distribution (i.e., the empirical activation probability) across states for each of these patterns. In these figures, we observe that:

(i) pattern (000), corresponding to the bias term, tends to model middle values for the percentage of votes for the three candidates (with an average percentage of votes of $\sim 50\%$ for the democrat candidate, $\sim 48\%$ for the republican candidate and $\sim 27\%$ for Perot), and activates mainly in the east and west coasts of the country, as well as Florida;

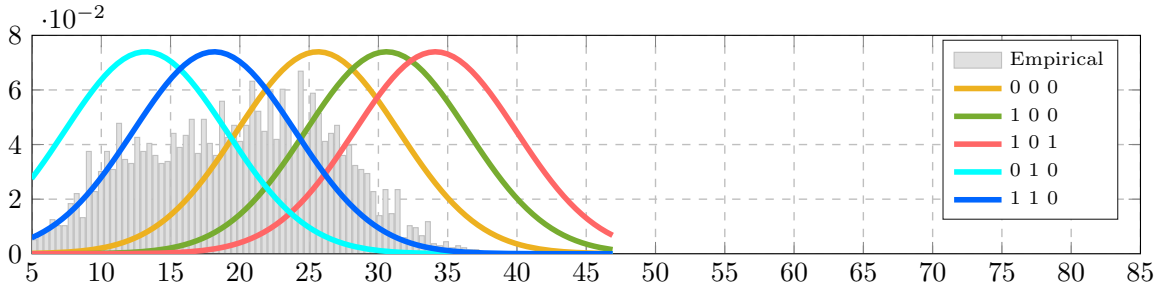
(ii) pattern (100) provides similar percentage of votes for the democrat and republican candidates as in pattern (000), but it favors the independent candidate Perot (with an average percentage of votes above 30%), and activates mostly in the north central-east region of the country and Maine (state where Perot's party managed to beat the Republican party);

(iii) pattern (101) activates in the north central-west region of the USA (not including the coast)

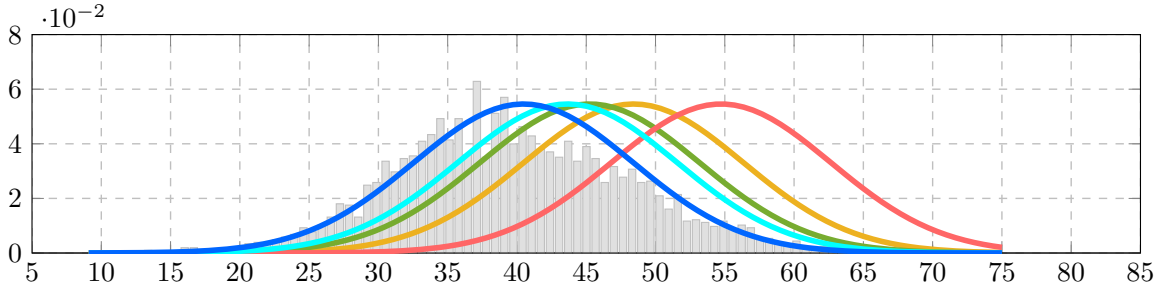
and represents a profile inclined towards the republican party (with an average percentage of votes of $\sim 55\%$) while also favoring in a lower extent the independent candidate; and

(iv) patterns (010) and (001) clearly capture democrat-oriented profiles, and activate mainly in the south east region of the USA, including the state from which Bill Clinton comes from, Arkansas.

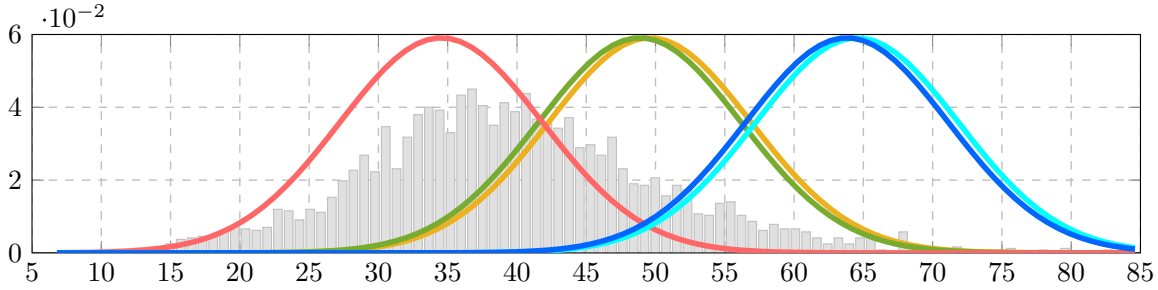
Note that the demographic results above are in agreement with the outcome of the election per counties,[13] as shown in Figure 4.9. Next, we analyze the demographic information associated to each of the feature patterns above. To this end, we show in Figure 4.7 the distribution of each attribute/dimension of the data for each of the considered patterns.



(a) % of votes cast for Ross Perot



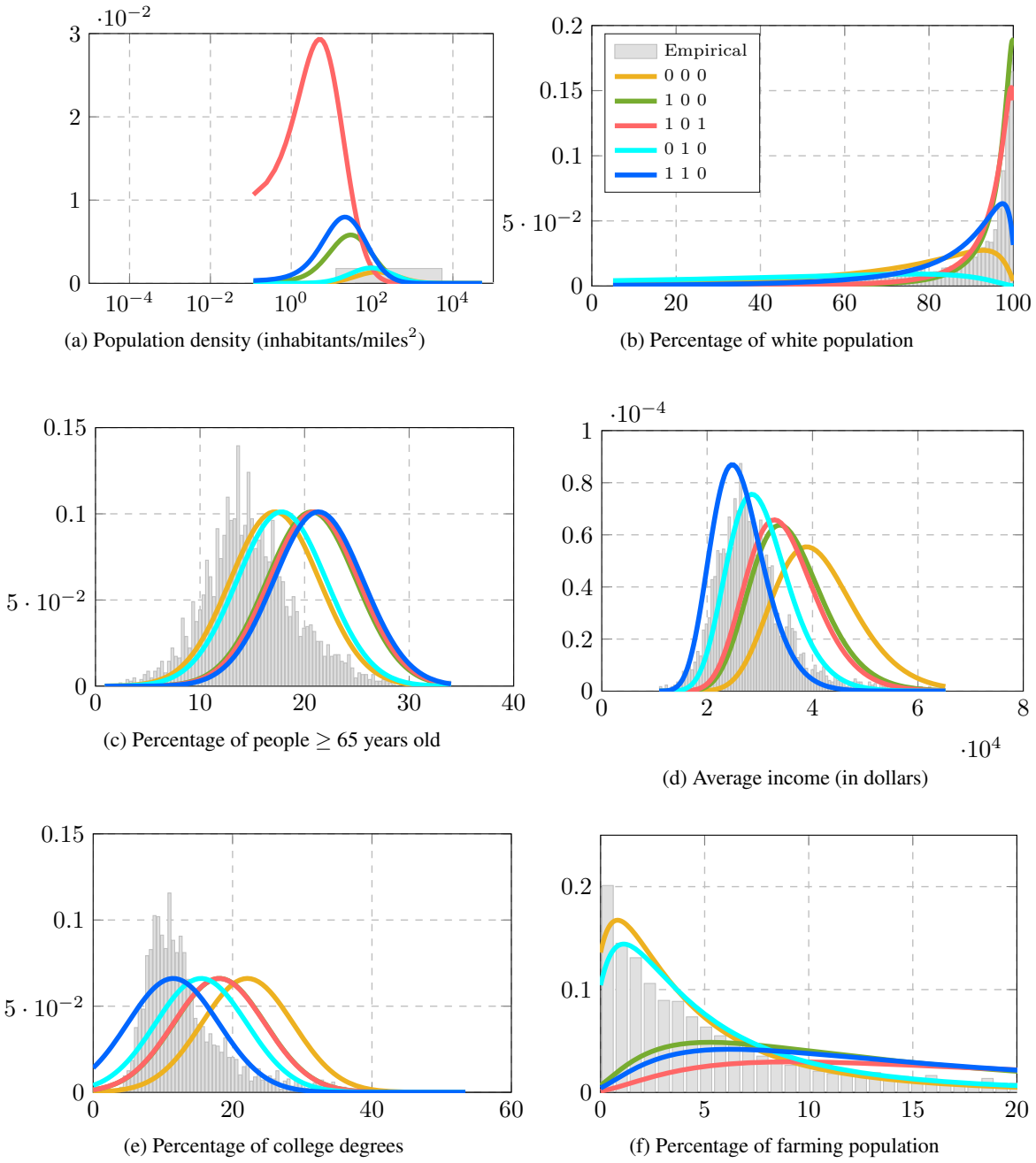
(b) % of votes cast for republican president



(c) % of votes cast for democrat president

Figure 4.6: **Inferred probability distribution for the five most popular patterns.** The patterns are sorted in the legend according to their degree of popularity, as described in Table 4.6.

[13] https://en.wikipedia.org/wiki/United_States_presidential_election,_1992



(a) Population density (inhabitants/miles^2)

(b) Percentage of white population

(c) Percentage of people $\geq$ 65 years old

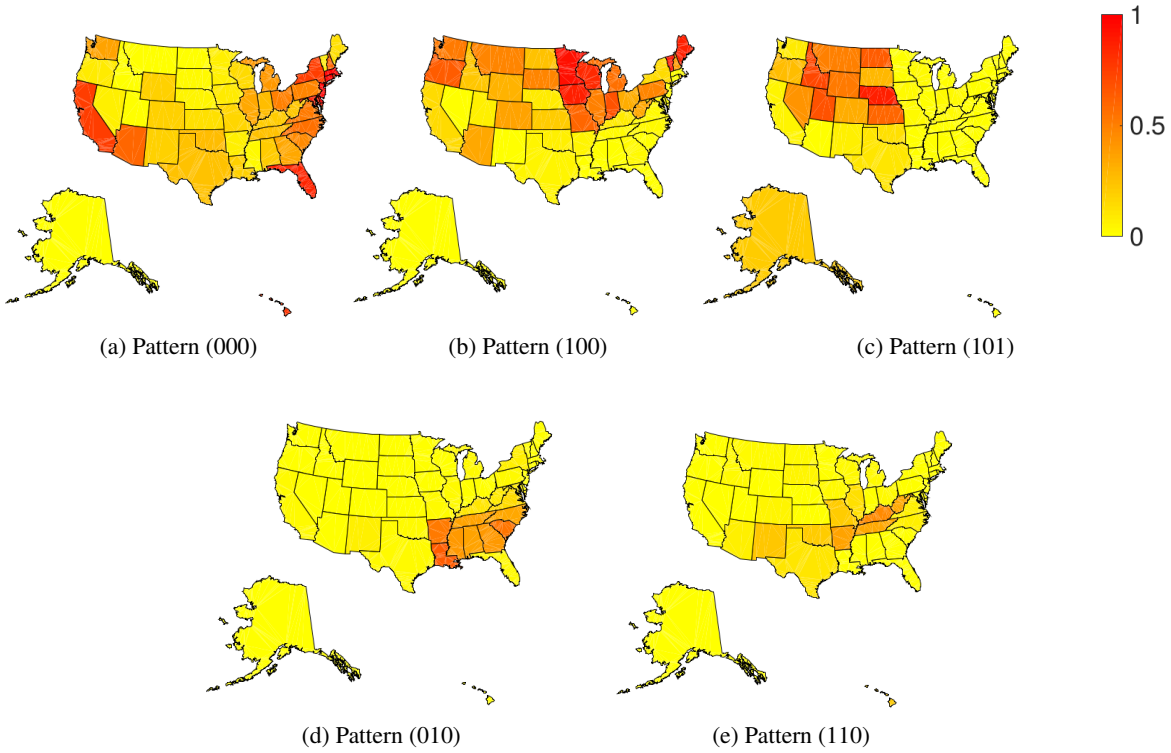
(d) Average income (in dollars)

(e) Percentage of college degrees

(f) Percentage of farming population

Figure 4.7: **Inferred probability distribution for the most occuring patterns.**

First, we observe that pattern (000), which activates mostly in the coasts and Florida, corresponds to the highest population density, average income, and percentage of college degrees, as well as an important race diversity and low farming activity. These observations align with the typical profile characterizing "big-cities". As stated before, this pattern is the most balanced in terms of voting tendency, with an equilibrated support for both democrat and republican, as well as intermediate



(a) Pattern (000) (b) Pattern (100) (c) Pattern (101)

(d) Pattern (010) (e) Pattern (110)

Figure 4.8: **Empirical probability of pattern activation per state.** We focus on the top-five most popular combinations of features. The label for each pattern indicates whether F1, F2, and F3 are active (value '1') or not (value '0'). Features F4 are F5 are always inactive in the five most common patterns, and thus are omitted in the labels.

values for the percentage of votes cast for Perot.

Second, patterns (100) and (101) represent the largest share of Perot's votes, both with an average percentage of votes above 30% for Perot. Figure 4.7 shows that Perot's main supporters, characterized mainly by pattern (101), also correspond to republican main supporters, who tend to live in low populated areas in the north central part of the country where farming activity is considerable, and the percentages of white population and over-65 years old population are also high. The second voting force backing Perot, captured by pattern (100) and located in the north east-central part of USA, corresponds mostly to white population with an intermediate-high average income and an average percentage of college degrees around 18% (the red curve in Figure 4.7e overlaps the green line). These results back the analysis in [115], which showed that the majority of Perot's voters (57%) were middle class, earning between $15,000 and $49,000 annually, with the bulk of the remainder drawing from the upper middle class (29% earning more than $50,000 annually). Perot's campaign ended up taking 18.9% of the votes, finishing second in Maine and Utah, as captured by pattern (100) and (101) respectively.

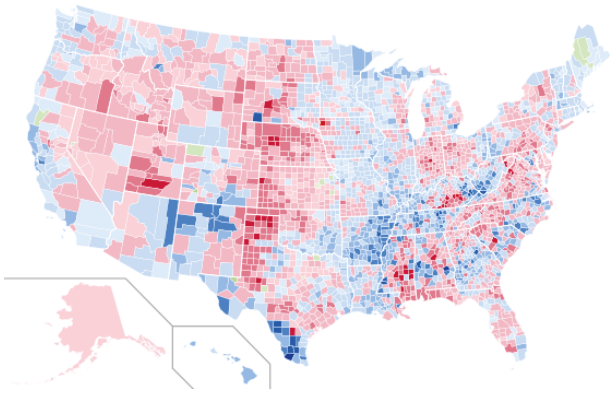


Figure 4.9: **Outcome of the 1992 presidential election per counties.** Blue color corresponds to a majority of votes for the democrat party, red corresponds to a victory for the republican party, green corresponds to a victory of the independent party of Ross Perot.

Finally, Democrat's patterns (010) and (110) are mainly active in the Southeastern United States, and capture a diverse range of voters in terms of their demographic properties. On the one hand, pattern (010) captures highly populated counties, with low values of family income, percentage of college degrees, percentage of white population and percentage of farming population. On the other hand, pattern (110) captures low populated counties with a large percentage of population above 65 year old, as well as a larger presence of farming activity and lower average income. These result might be explained by the broad appeal across all socio-ethno-economic demographics that the Democratic party has historically targeted.

4.B Appendix: Details on the phase II Clinical Trial for Codrituzumab

4.B.1 List of Biomarkers

Biomarker	Description	Type of Biomarker
Age	Age	Clinical data and demographics
F	Sex Female	Clinical data and demographics
AS	Race Asian	Clinical data and demographics
AA	Race African American	Clinical data and demographics
OT	Race Other	Clinical data and demographics
WH	Race White	Clinical data and demographics
Weight	Baseline Weight	Clinical data and demographics
Height	Baseline Height	Clinical data and demographics
BMI	Baseline Body Mass Index (kg/m^2)	Clinical data and demographics
D	Duration of Exposure (days)	Clinical data and demographics
ECOG	ECOG Performance Status at Baseline	Clinical data and demographics
Sorafenib	Prior Sorafenib Treatment	Clinical data and demographics
CP score	Child Pugh Score	Clinical data and demographics
SDTL	Sum diameter (measurable) target lesions	Clinical data and demographics
IHC 0	GPC3 IHC	Tumor Histology
IHC 1+	GPC3 IHC 1+	Tumor Histology
IHC 2+	GPC3 IHC 2+	Tumor Histology
IHC 3+	GPC3 IHC 3+	Tumor Histology
Vascular Invasion	Macrovascual Invasion or Extrahepatic	Tumor Histology
CD16	CD16 cell density	Tumor Histology
CD3	CD3 cell density	Tumor Histology
%CD16	% OF CD16 in stroma	Tumor Histology
%CD3	% OF CD3 in stroma	Tumor Histology
%Stroma	% of stroma	Tumor Histology
CD3/CD16 necrotic	CD3/CD16 count in necrotic tissue	Tumor Histology
CD3/CD16 tumor	CD3/CD16 count in tumor	Tumor Histology
CD3/CD16 viable	CD3/CD16 count in viable cells	Tumor Histology
%Necrotic	% necrotic cells in tissue	Tumor Histology
%Tumor	% tumor in tissue	Tumor Histology
%Viable	% viable cells in tissue	Tumor Histology
H score Cyt	GPC3 H score cytoplasmic	Tumor Histology
H score Mem	GPC3 H score membrane	Tumor Histology
Ctrough	Ctrough at Cycle3 Day1	Exposure
AFP	Alpha Fetoprotein	Circulating protein
AAT	Alanine Aminotransferase	Circulating protein
CRP	C Reactive Protein	Circulating protein
sGPC3 114/165	Soluble GPC3 GT114-GT165	Circulating protein
sGPC3 30/57	Soluble GPC3 GT30-GT57	Circulating protein
sGPC3 30/607	Soluble GPC3 GT30-GT607	Circulating protein
sGPC3 11/96	Soluble GPC3 M3C11-GT96	Circulating protein

ADCC CD107	Antibody-dependent cell cytotoxicity CD107A activity	Blood cell activity
ADCC CD16	Antibody-dependent cell cytotoxicity CD16 activity	Blood cell activity
CD45	CD45+ cell count	Blood cell subset
B	B cell count	Blood cell subset
CD3	CD3+ cell count	Blood cell subset
CD4	CD4+ cell count	Blood cell subset
CD8	CD8+ cell count	Blood cell subset
CD4/CD8	CD4/CD8 ratio	Blood cell subset
CD8 NK	CD8+ NK cells	Blood cell subset
CD16	CD16+ cell count	Blood cell subset
CD56-CD16+	CD56-CD16+ NK cells	Blood cell subset
CD56bright	CD56bright NK cells	Blood cell subset
CD56dimCD16-	CD56dimCD16- NK cells	Blood cell subset
CD56dimCD16bright	CD56dimCD16bright NK cells	Blood cell subset
NK	NK cell count	Blood cell subset
CD56	CD56+ cell count	Blood cell subset
NKP46	NKP46+ NK cell count	Blood cell subset
DN	Double negative cells	Blood cell subset
DP	Double positive cells	Blood cell subset
CD16MESF	CD16 MESF	Blood cell subset
NKP46MESF	NKP46 MESF	Blood cell subset
FCGRIIIA-158_A/A	FCGRIIIA-158 A/A	DNA polymorphism
FCGRIIIA-158_C/A	FCGRIIIA-158 C/A	DNA polymorphism
FCGRIIIA-158_C/C	FCGRIIIA-158 C/C	DNA polymorphism
FCGRIIIA-158_NA	FCGRIIIA-158 NA	DNA polymorphism
FCGRIIA-131_A/A	FCGRIIA-131 A/A	DNA polymorphism
FCGRIIA-131_A/G	FCGRIIA-131 A/G	DNA polymorphism
FCGRIIA-131_G/G	FCGRIIA-131 G/G	DNA polymorphism
FCGRIIA-131_NA	FCGRIIA-131 NA	DNA polymorphism
PFS	Progression Free Survival (Months)	Clinical endpoint

4.B.2 Study Design and Patients

Adult patients with unresectable advanced or metastatic HCC who were previously treated with at least one line of systemic agent and with progressive disease were enrolled in a randomized, placebo-controlled, double-blind, multicenter phase II trial (NCT01507168). Patients received either intravenous Codrituzumab at 1600 mg every two weeks or placebo (with a patient ratio of treatment:placebo of 2:1) until disease progression, and were followed for overall survival. Details of study design have been previously described [2].

4.B.3 GPC3 Expression in Tumor

All patients enrolled in the study provided a tumor tissue sample to determine the level of GPC3 expression by immunohistochemistry (IHC) under central review prior to study entry[2]. Specifically, IHC was performed based on a formalin-fixed, paraffin-embedded (FFPE) block of the primary tumor or the metastatic site collected within approximately 12 months prior to informed consent or four 4μm thick unstained slides (freshly cut from a FFPE block of the primary tumor or the metastatic site obtained within approximately 12 months prior to informed consent). If no archival material was available, a pre-treatment core needle biopsy with minimally 18-gauge needle was obtained. The IHC staining was done on BenchMark XT (Ventana Medical Systems, Inc. or VMSI, catalog number 750-700) or ULTRA (VSMI, catalog number 750-600) platforms. Each patient was assigned a GPC3 IHC score with ordered categorical values 0, 1+, 2+, and 3+, corresponding to increasing levels of GPC3 expression, with scores 0 and 3+ indicating the lowest and highest levels of GPC3 expression, respectively.

4.B.4 Flow Cytometry

Surface cell markers from circulating blood cells were measured by flow cytometry. Lymphocyte subsets were assayed using Trucount tubes (Becton, Dickinson and Company or BD, catalog number 340334). The expression level of CD16 on NK cells was measured by flow cytometry analysis of the pre-treatment peripheral blood mononuclear cells using CD16-specific antibody. Measurement was done on FACSCantoTM II (BD, catalog number 657338). The quantification of CD16 expression level, or fluorescence intensity in units of Molecules of Equivalent Soluble Fluorochrome (MESF), denoted by CD16 MESF, was calculated by converting fluorescence measurements of the NK cell population to an MESF value based on an MESF calibration curve prepared according to fluorescence intensity of calibration beads (QuantumTM MESF bead standard, manufactured by Bang Laboratories, IN, USA) [116].

4.B.5 Soluble Protein Measurements

Monoclonal antibodies against soluble GPC3 protein were generated as previously described12. Five anti-N-terminal fragment mAbs (designated GT30, GT95, GT114, GT165 and GT607), and two anti-C-terminal fragment mAbs (designated GT57 and M3C11), were used in combination in four different assays (sGPC3 114/165, sGPC3 30/57, sGPC3 30/607, sGPC3 11/96) to detect full length GPC3 protein, or any possible cleavage fragments containing N- or C-terminus. The protocol for

sandwich ELISA assay has been described [81].

4.B.6 DNA Polymorphisms

Genomic deoxyribonucleic acid (DNA) was extracted from blood samples by using a QiAmp Blood Mini Kit (Qiagen, Germany). DNA concentrations were measured by using NanoDrop ND-1000 (Thermo Fisher Scientific, Wilmington, DE, USA), and DNA samples were diluted in nuclease free water to get a final concentration of 1 ng/μl. Patients were genotyped for two different Fc gamma receptor polymorphisms, FcgRIIa-H131R and FcgRIIIa-V158F using TaqMan technology on Applied Biosystems (AB) 7500 Fast Real-Time PCR system (Applied Biosystems Inc., CA, USA). Probes and primers (TaqMan SNP Assays for rs1801274 and rs396991) were ordered from Applied Biosystems. Genotyping was performed following the manufacturer's instructions.

4.B.7 Further Results

The mentioned subpopulations in the following table correspond to those listed in Table 4.1.

Continuous variables	Effect size	Mw-test
AFP	3.90	**4.71e-06**
H score Mem	1.34	**9.64e-15**
H score Cyt	0.75	**4.66e-11**
sGPC3 30/607	0.31	**1.62e-06**
sGPC3 114/165	0.29	**1.26e-06**
CD4	0.10	**7.74e-05**
CD3	0.09	**9.16e-06**
CD45	0.07	**4.04e-05**
sGPC3 11/96	0.48	**9.40e-04**
CD56-CD16+	0.40	**1.09e-03**
sGPC3 30/57	0.39	**5.42e-04**
CD56dimCD16-	0.22	**1.81e-03**
CD8	0.11	**8.52e-04**
CD56dimCD16bright	0.47	**6.50e-03**
NK	0.46	**4.13e-03**
CD56	0.17	**6.95e-03**
DN	0.17	**1.78e-02**
NKP46	0.11	**3.13e-03**
CD16	0.11	**4.19e-03**
Discrete variables	**Effect size**	**F-test**
IHC 3+	0.82	**1.13e-20**
IHC 1+	-0.37	**2.29e-06**
IHC 2+	-0.24	**1.03e-04**
IHC 0	-0.21	**3.49e-04**

(a) both F1 and F2 inactive

Continuous variables	Effect size	Mw-test
P Necrotic / stroma	1.92	**4.38e-03**
NK	1.10	**3.85e-03**
CD56dimCD16bright	1.02	**7.74e-03**
CD56-CD16+	0.60	**7.88e-03**
CD56dimCD16-	0.56	**6.37e-03**
CD16	0.19	**5.18e-03**
P Tumor Viable	-0.06	**4.38e-03**
Discrete variables	**Effect size**	**F-test**
-	-	-

(b) F1 active and F2 inactive

Table 4.8: **Significant biomarkers associated with better prognosis in the treatment arm for different subpopulations.** (a) Comparison of subpopulation 6 vs 5, (b) comparison of subpopulation 10 vs 9. Continuous variables on the top, discrete variables on the bottom. Colors code for the different statistical significance levels: $\alpha_s = 0.001$ (black), $\alpha_s = 0.005$ (blue), and $\alpha_s = 0.01$ (green).

5

Hierarchical Indian Buffet Process for Discovery of Genetic Associations

Personalized medicine aims at combining genetic, clinical, and environmental data to improve medical diagnosis and disease treatment, tailored to each individual [32]. Within this framework, this chapter presents a Bayesian nonparametric (BNP) approach to identify genetic associations with clinical features in cancer, in order to enhance clinical diagnosis and further understanding of the disease. We propose a hierarchical approach, the hierarchical Poisson factor analysis (H-PFA) model, to share information across patients having different types of cancer. To discover statistically significant associations, we follow a similar statistical procedure to the one presented in Chapter 4, which combines Bayesian modeling with bootstrapping techniques, and corrects for multiple hypothesis testing. We compare the results of H-PFA with two other classical methods in the field: case-control (CC) setups, and linear mixed models (LMMs).

5.1 Introduction

Cancer encompasses not one, but a vast group of genetic diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. Although a small set of common underlying principles have been identified (the so-called "hallmarks" of cancer [79, 80]), each type of cancer presents very unique properties, making this disease very hard to handle [151, 85, 194]. In the last decade, high-throughput genotyping technologies have led to the discovery of cancer-correlated gene mutations, most of which were not previously suspected to be related to carcinogenesis [49]. However, only the gene mutations with very strong effects have been discovered and many other genes with weaker effects still remain to be found [10]. Genetic-association studies have been widely used in the search for such genes, but success has been limited so far [151].

A first difficulty in cancer association studies is the immense phenotypic heterogeneity, which reduces statistical power in the discovery method and causes some associations to remain hidden [164, 110, 176]. Second, cohort sizes tend to be small, especially in rare cancers, which makes the discovery of small effect size associations difficult [10]. Third and last, cancers are driven by the accumulation of mutations that may act epistatically or pleiotropically during the course of the disease [213, 36, 183]. *Epistasis* refers to complex interactions between genetic variants that have an effect on the same phenotype, while *pleiotropy* means that multiple phenotypes are influenced by the same single mutation. Indeed, cancer is known to be polygenic and present complex pleiotropic phenotypes [49]. New approaches need to be found in order to overcome these difficulties.

In recent years, efforts to mine electronic health records (EHRs) show promise to impact nearly every aspect of healthcare [101]. The adoption of EHRs in hospitals has increased dramatically, and has become an interesting resource for phenotyping [4, 129], with the potential for establishing new patient-stratification principles and for revealing unknown disease correlations. Integrating EHRs data with genetic data will also give a finer understanding of genotype-phenotype relationships [214]. EHRs consist of both structured and unstructured information. Structured data is a valuable source of information that includes billing codes, laboratory reports, physiological measurements, and demographic information, among others. Yet, most of the clinical data comes as unstructured notes, e.g., around 98% of the EHRs [101]. These include a broad spectrum of clinically-relevant information which might be useful to identify novel phenotypic relationships so far unknown by the clinicians [129, 101].

This chapter presents a joint generative model to discover associations between genetic mutations and clinical features in cancer that deals with phenotype heterogeneity, small cohort size, epistasis and pleiotropy in a straightforward way. Our method is a generative model that infers latent topics

from the clinical text, associated to hidden genetic factors that directly capture complex interactions between the genes. It is directly inspired on the Poisson factorization model for recommendation systems in [69], with three important differences. First, we introduce confounding effects as conditional variables, i.e., variables that might cause spurious associations to appear. In particular, our model considers multiple types of cancer together (the type of cancer is treated as a confounder), and shares information among all patients in a hierarchical fashion. Indeed, most cancers are known to share a common pathogenesis despite specificities of the cell type and tissue origin [194]. By doing so, specific effects for each type of cancer can be isolated, and novel (less well-known) associations with gene mutations of smaller effect size can be obtained. Second, we force sparsity on the textual and genetic topics by using shape parameters in the Gamma distribution priors smaller than one. Third, we present a nonparametric alternative model to the work in [69] by replacing the continuous patient weights with a binary matrix whose probability distribution is induced by a hierarchical extension of the Indian buffet process (IBP). Also in the literature, the authors in [70] propose a nonparametric Poisson factorization model, but they rely on a stick-breaking construction different from the IBP, and the weights are continuous, which renders interpretability of the latent variables more tedious. The discrete nature of the IBP helps in terms of interpretability, and allows combining the proposed Bayesian model with classical frequentist approaches for statistical testing between the inferred patient partitions. An efficient Markov chain Monte Carlo (MCMC) procedure based on a slice sampler for the hierarchical IBP is presented.

Bayesian modeling has already been proven useful for epistasis [225], pleiotropy [225, 224] or sub-phenotyping aplications [147, 107]. Up to our knowledge, the proposed model is the first one to deal with clinical text data and genetic information jointly, capturing phenotypic heterogeneity, epistasis and pleiotropy in a straightforward way while correcting for the cancer type as confounder. We consider multiple cancers jointly in order to increase statistical power, allow for the analysis of rare cancers, and identify fundamental mechanisms shared across different types of cancer.

5.2 Genetic Association Studies

Any two human genomes differ in millions of different ways. Variations include individual nucleotides of the genomes as well as larger variations, such as deletions, insertions and copy number variations. Any of these variations may cause alterations in an individual's traits, or *phenotype*, which can be anything from disease risk to physical properties such as eye color or height [193]. We distinguish between two types of variations: *somatic* mutations and *germline* mutations. A somatic (also

called "acquired") mutation is a change in the genetic structure that is not inherited from a parent, nor passed to offsprings: these are not inherited because they do not affect the reproductive cells (sperm and egg) but any other type of cell. In contrast, germline mutations affect the reproductive cells, and can thus be inherited. In this thesis, we solely focus on somatic mutations, which usually appear by environmental causes, such as ultraviolet radiation or any exposure to certain chemicals. Genetic association studies arise as a way to examinate a genome-wide set of genetic variants in different individuals to see whether any variant is associated with a phenotypic trait. Such studies have become popular in the last decades in order to study the relationship between the genotype and phenotype of individuals [89, 114] and, in particular, in cancer research [49, 148].

Earlier genetic studies were focused on the effect of individual single-nucleotide polymorphisms (SNPs), and are referred to as genome-wide association study (GWAS). We call SNP to a variation in a single nucleotide that occurs at a specific position in the genome: these are the most common type of variation among individuals [122]. For example, at a specific base position in the human genome, the base C may appear in most individuals, but in a minority of subjects, the position is occupied by base A. We then say that there is a SNP at this specific base position, and the two possible nucleotide variations – C or A – are said to be *alleles* for this base position. A GWAS is a statistical analysis which aims at finding individual correlations between any SNP and a particular phenotype [114].

Since the beginning of genetic association studies, there have been two general trends in the literature. On the one hand, a wide range of different phenotypes have been considered, including protein-protein network, gene expression, biomarkers, or intermediate clinical phenotypes at the organism level [151]. On the other hand, bigger sample size scenarios have been analyzed, with studies of up to 200,000 individuals. Despite the efforts to gather information for many individuals, privacy and ethical issues remain an important drawback for this task, such that most genetic association studies still work in the order of thousands of individuals.

In addition to the screening of genetic associations, it is also common to take into account any variables that could potentially confound the results. A *confounder* or confounding factor is a variable that, if not observed, causes two other variables to appear correlated, while in fact these two are independent given the confounder. A typical confounder example would be smoking to "drinking coffee" and "having lung cancer". Gender and age are other common examples of confounding variables. Moreover, it is also known that many genetic variations are associated with the geographical and historical populations in which the mutations first arose.

Lack of well defined case and control groups, insufficient sample size, control for multiple testing, and control for population stratification are common problems in genetic association stud-

ies [114, 10]. Particularly the statistical issue of multiple testing is problematic as "the massive number of statistical tests performed presents an unprecedented potential for false-positive results" [150]. In our analysis, we resort to either the Bonferroni correction or the Benjamini-Hochberg procedure [17] to control for multiple hypothesis testing, the former controls for the family-wise error rate (FWER) (probability to reject at least one true hypothesis) while the latter estimates the false discovery rate (FDR) and thus controls for the number of false positives.

5.2.1 Standard Approach: Case-Control Setup

The most common approach of GWAS is the case-control (CC) setup, which compares two large groups of individuals, one healthy control group and one case group affected by a disease [31, 122]. All individuals in each group are genotyped for the majority of common known SNPs. For each of these SNPs, it is then investigated if the allele[1] frequency is significantly altered between the case and the control group. In such setups, the fundamental unit for reporting effect sizes is the odds ratio. In the context of GWAS, the odds ratio refer to the odds of disease for individuals having a specific allele and the odds of disease for individuals who do not have that same allele. When the allele frequency in the case group is much higher than in the control group, the odds ratio is higher than one, and vice versa for lower allele frequency. A p-value for the significance of the odds ratio is typically computed using a simple χ-squared test or Fisher test. Finding odds ratios that are significantly different from one is the objective of the GWAS because this shows that a SNP is associated with the disease.

5.2.2 Confounder Correcting Approach: Linear Mixed Model

linear mixed models (LMMs) have proved particularly useful for GWAS due to its capacity to account for confounding effects and limit the number of false associations [121, 120]. Let x_{ng} be the indicator variable for a somatic mutation in gene $g \in \{1, \ldots, G\}$ for a particular patient $n \in \{1, \ldots, N\}$. The variable x_{nd} is binary and indicates whether any somatic mutation occurred in the corresponding gene. Let y_{nq} be the binary indicator variable of the presence of a certain clinical feature $q \in \{1, \ldots, Q\}$ for a given patient n. Finally, let us define c_{nl} as the binary assignment variable of patient n to the cancer type $\ell \in \{1, \ldots, L\}$, where $\sum_\ell c_{n\ell} = 1$ (we only consider patients having one single type of cancer). For each pair of gene g and clinical feature q, a LMM can be defined as follows:

$$\boldsymbol{y}_{\bullet q} = \boldsymbol{x}_{\bullet g}\beta_{qg} + \boldsymbol{u}_{\bullet qg} + \boldsymbol{\varepsilon}_{\bullet qg}, \tag{5.1}$$

[1] An *allele* refers to each of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome.

where $\beta_{qg} \in \mathbb{R}$ refer to the fixed effect, and $\boldsymbol{u}_{\bullet qg}$, $\boldsymbol{\varepsilon}_{\bullet qg} \in \mathbb{R}^{N\times 1}$ are the random effects (structured noise and observational noise, respectively). The prior assumptions for the structured and uniform noises are $\boldsymbol{u}_{\bullet qg} \sim \mathcal{N}(0, \sigma_u^2\mathbf{K})$ and $\boldsymbol{\varepsilon}_{\bullet qg} \sim \mathcal{N}(0, \sigma_e^2\mathbf{I})$, where $\mathbf{K}$ refers to a similarity matrix between the patients, for instance, the cosine similarity of the cancer type assignment vectors $\mathbf{c}_{i\bullet}$ and $\mathbf{c}_{j\bullet}$, $\mathbf{K} = \boldsymbol{C}\boldsymbol{C}^T$. The LMM assumes that the output $\boldsymbol{y}_{\bullet \boldsymbol{q}}$ is Gaussian-distributed:

$$\boldsymbol{y}_{\bullet q} \sim \mathcal{N}(\boldsymbol{x}_{\bullet g}\boldsymbol{\beta}_{qg}, \sigma_u^2\mathbf{K} + \sigma_e^2\mathbf{I}). \tag{5.2}$$

When the data is binary or count data, a common practice is to apply a standard rank-based inverse normal transformation beforehand as a preprocessing step [96].

5.3 Our Approach

Let $\mathbf{X} \in \mathbb{N}^{N\times D}$ be the observation matrix of count data for N patients and D dimensions, where D includes both clinical and genetic information, i.e., $D = G + Q$, where G is the number of genes and Q is the number of clinical terms. In the following, we propose two Poisson factor analysis (PFA) approaches to model the joint observation matrix $\mathbf{X}$ of genetic information and clinical data. In these models, patients will be represented by binary feature activation vectors, and each of these features will capture common correlation patterns among the somatic mutations and clinical term occurrences.

5.3.1 Bernoulli Process Poisson Factor Analysis

We first consider a simple nonparametric non-negative matrix factorization model with Poisson likelihood and Gamma-distributed factors:

$$x_{nd} \sim \text{Poisson}\big(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}\big), \tag{5.3}$$

$$\mathbf{B}_{kd} \sim \text{Gamma}\Big(\alpha_{\mathbf{B}}, \frac{\mu_{\mathbf{B}}}{\alpha_{\mathbf{B}}}\Big), \tag{5.4}$$

$$\mathbf{Z} \sim \text{IBP}(\alpha), \tag{5.5}$$

where α is the concentration parameter of the IBP controlling the a priori number of ones in matrix $\mathbf{Z}$ (i.e., the a priori expected number of latent features), and $\mu_{\mathbf{B}}$, $\alpha_{\mathbf{B}}$ are the prior mean and shape parameter for each element of matrix $\mathbf{B}$. Sparsity can be induced easily in the factors by choosing $\alpha_{\mathbf{B}} \ll 1$. From now on, we will refer to this basic model as the Bernoulli process Poisson factor analysis (BeP-PFA). The "Bernoulli process (BeP)" designation in the naming emphasizes that each

feature activation vector $\mathbf{Z}_{n\bullet}$ is a draw from a BeP.[2] Inference is performed using a MCMC approach based on a semi-ordered stick-breaking representation of the IBP prior [202]. A complete description of the inference algorithm can be found in Section 8.2.2 of the final appendix.

5.3.2 Hierarchical Bernoulli Process Poisson Factor Analysis

Although different types of cancer are known to share similar phenotypes and underlying mechanisms (shared activation of certain pathways), the mutation rate and phenotype occurrence might vary in different proportions, according to each type of cancer. Given this premise, we propose a hierarchical Bernoulli process Poisson factor analysis model to allow for different feature activation levels depending on each type of cancer. In the following, we will shorten the name of this model to hierarchical Poisson factor analysis (H-PFA).

Let $r_n \in [1, \dots, L]$ be a categorical variable indicating the type of cancer of patient n among the total number of cancer types L (in the previous notation from Section 5.2, r_n corresponds to the index of the non-zero value in vector $\mathbf{c}_{n\bullet}$). A hierarchical construction can be formulated based on the finite representation of the IBP and letting $K \to \infty$, such that different levels of feature activation are allowed for each type of cancer. Let ρ_k be the global activation probability of feature k, and π_k^ℓ be the specific activation probability of feature k for the cancer type $\ell \in [1, \dots, L]$. We can then assume that each specific activation probability is Beta-distributed such that $\mathbb{E}_\ell[\pi_k^\ell] = \rho_k$:

$$\begin{aligned} \rho_k &\sim \text{Beta}\Big(\frac{\alpha}{K}, 1\Big) & z_{nk} &\sim \text{Bernoulli}(\pi_k^{r_n}) \\ \pi_k^\ell | \rho_k &\sim \text{Beta}\Big(\frac{\rho_k}{1-\rho_k}, 1\Big) & \mathbf{B}_{kd} &\sim \text{Gamma}\Big(\alpha_{\mathbf{B}}, \frac{\mu_{\mathbf{B}}}{\alpha_{\mathbf{B}}}\Big) \end{aligned} \tag{5.6}$$

$$\mathbf{x}_{nd} \sim \text{Poisson}\big(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}\big), \tag{5.7}$$

where the feature activation variables in vector $\mathbf{Z}_{n\bullet}$ are drawn from different activation probability vectors $\{\boldsymbol{\pi}^1, \dots, \boldsymbol{\pi}^L\}$ depending on the type of cancer c_n of patient n. When $K \to \infty$, this prior over $\mathbf{Z}$ is equivalent to a hierarchical Beta process (BP) construction [206] on top of BePs in the De Finetti representation introduced in Section 2.3.2. In the same way that a hierarchical Dirichlet process (HDP) allows for atom sharing with varying weights across different groups of data (see Section 2.3.1), the H-PFA allows for feature sharing with different activation weights across different types of cancer.

[2] Traditionally, PFA refers to both continuous weights and dictionary components [226].

5.4 Results

5.4.1 Database Description

So far, genomic testing of tumors has been done routinely only for certain solid cancer tumors, such as melanoma, lung, or colon cancer. For most cancers, the available tests have been limited to analyzing one or a handful of genes at a time, and within each gene, only the most common mutations could be detected.[3] A new targeted tumor sequencing test called MSK-IMPACT^{TM} (Integrated Mutation Profiling of Actionable Cancer Targets) is able to detect gene mutations and other critical genetic aberrations in both rare and common cancers [29].

Using the MSK-IMPACT technique, somatic mutations regarding specific screened genes can be obtained as follows. For each patient, tumor cells are compared with healthy cells of that same patient, extracted from the blood stream. In this chapter, a gene is said to be mutated when there exists at least one difference in the sequence between the tumor cells and healthy cells for that particular gene.[4] We finally obtain a binary matrix for $N = 1946$ patients and $G = 410$ genes where "1" encodes a mutated gene and "0" otherwise. The screened genes have been shown to play a role in the development or behavior of tumors,[5] although their individual relation to specific phenotypes remains obscure [29].

Concerning the clinical information, based on all EHRs, we build a bag-of-word representation of unified medical language system (UMLS) terms, extracted using the *Metamap*[6] processing tool [14]. The UMLS refers to a standardized, comprehensive thesaurus and ontology for biomedical concepts, whose objective is to provide facilities for natural signal processing tasks [22]. Since each patient can have a varying number of EHRs, we group all clinical history into a single EHR, and only consider the appearance or absence of each UMLS term. We compute the tf-idf score for each UMLS term, and only keep the 300 clinical terms with highest score.

The final database includes clinical and genetic information for $N = 1946$ patients and 5 different cancer types: bladder cancer, breast carcinoma, colorectal cancer, non-small cell lung cancer, and prostate cancer. We consider genes and UMLS terms that are present in at least 1% of the patient population, resulting in $D = 249$ dimensions, including 72 genes and 177 clinical terms. Even if the dataset is binary, we can use a Poisson likelihood because of the high sparsity degree of such matrix (7.28% of non-zero values). In such scenario, the Poisson distribution is a good approximation of a

[3] `https://www.mskcc.org/msk-impact`

[4] The considered sequencing technology is able to remove most of the technical noise, in contrast to other technologies.

[5] `https://www.mskcc.org/blog/new-tumor-sequencing-test-will-bring-personalized-treatment-options-more-patients`

[6] Source code available at: `https://metamap.nlm.nih.gov/`

Bernoulli, and we adopt it by mathematical convenience in the inference process.

5.4.2 Experimental Setup

We compare the proposed H-PFA approach with a LMM and a standard case-control set-up for each potential clinico-genetic association. The model parameters for each LMM are found by maximizing the log likelihood using standard optimization techniques within a python platform called LIMIX [120]. In the final step, we obtain p-values for each pair $(y_{\bullet q}, x_{\bullet g})$ using likelihood ratio tests. Regarding the case-control analysis, for each clinical term we consider a case and control group corresponding to the patients having that clinical term active or inactive respectively. Given such partition, we perform an individual Fisher test for each gene. For all methods, we correct for multiple hypothesis testing based on the Benjamini-Hochberg approach [17]. Finally, concerning the H-PFA model, we follow the statistical procedure described in Section 4.3.3. As already described in the last chapter, we want to make our results more robust against outliers and reduce the number of false associations, so that we take multiple posterior samples and perform bootstrapping one hundred times of the whole process. For each statistical test, we obtain one hundred p-values, one for each bootstrapping instance, which are summarized via the median p-value.[7] An association is said to be significant when this median p-value is higher than the adjusted significance level α_s after correction for multiple hypothesis testing. For all simulations, we adopt a significance threshold value 0.001 and only consider associations with a positive effect size, for simplicity. In order to increase model interpretability, we also introduce one bias term to capture mean effects corresponding to cancer in general. This bias term is forced to be active for all patients. Finally, we have set the hyperparameters of the proposed H-PFA as $\alpha_{\mathbf{B}} = 0.01$ and $\mu_{\mathbf{B}} = 1$, while we infer the values for the concentration parameter α as described in Appendix 8.

5.4.3 Identification of Clinico-Genetic Associations

Figure 5.1 represents the number of associations found by each method, and how many overlap across techniques. LMM found 14 clinico-genetic associations, CC found 178, and H-PFA found 95.[8] As expected, LMM is the method that finds the least number of associations, since it corrects for the cancer type as a confounder effect, in order to get rid of cancer-specific associations and only get less well-known associations that are present across all types of cancer. CC is the method that

[7] Note that statistical significance could also be accounted for using Bayesian factors or posterior predictive checks [60]. We here adopt the most established approach for statistical significance assessment within this field.

[8] The H-PFA model is very flexible, as it can also find correlations between the genes, or between the clinical terms. 95 is the number of clinico-genetic associations only.

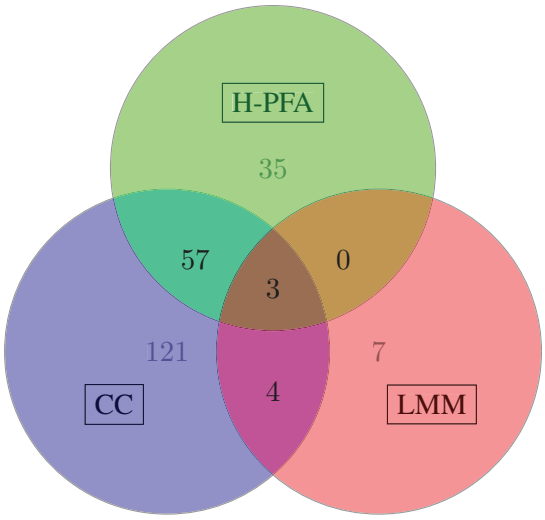


Figure 5.1: **Venn Diagram of number of associations.**

	Phenotype	Gene
ALL	Stage IV Lung Adenocarcinoma	EGFR
	Adenocarcinoma of lung (disorder)	EGFR
	Malignant neoplasm of urinary bladder	TERT
CC ∩ LMM	KRAS gene	KRAS
	Potassium Ion	KRAS
	Pulmonary function tests	STK11
	Rash and Dermatitis Adverse Event Associated with Chemoradiation	TCF7L2
CC ∩ H-PFA	Colorectal Carcinoma	APC
	capecitabine	APC
	Colorectal	KRAS
	Lumpectomy of breast	PIK3CA
	Adriamycin	PIK3CA
	Renal function	ARID1A
	Simple mastectomy	GATA3
	Hydronephrosis	KDM6A
	...	...

Figure 5.2: **Shared associations across methods.**

discovered the highest number of associations, from which only 30% are shared with H-PFA. Out of the 95 associations discovered by the H-PFA approach, 63% were also present in any of the other methods. Figure 5.2 lists the associations that are shared across methods. Tables 5.1 and 5.2 present the list of clinico-genetic associations found by LMM and CC methods (for CC, we only report a random selection of associations, but the complete list can be found in Appendix 5.A).

Phenotype	**Gene**	β_{qg}	p**-value**
pump (device)	APC	0.61	2.78e-53
S3 (sacral segmental innervation)	TP53	0.29	1.02e-09
Stage IV Lung Adenocarcinoma	EGFR	0.32	2.07e-29
Folinic Acid-Fluorouracil-Irinotecan Regimen	APC	0.59	1.08e-60
Folinic Acid-Fluorouracil-Irinotecan Regimen	KRAS	0.30	1.16e-18
Hepatectomy	APC	0.65	1.21e-52
Hepatectomy	KRAS	0.27	7.27e-12
FOLFOX Regimen	APC	0.66	5.12e-113
Tract	ARID1A	0.21	5.61e-12
Malignant neoplasm of urinary bladder	TERT	0.55	1.07e-61
Renal function	TERT	0.28	8.40e-12
Flushing	APC	0.30	1.54e-11
Non-Small Cell Lung Carcinoma	EGFR	0.18	4.11e-11
Colorectal Carcinoma	APC	0.61	2.34e-63
Adenocarcinoma of lung (disorder)	EGFR	0.26	1.49e-22
Simple mastectomy	PIK3CA	0.16	3.40e-08
Immunotherapy	TERT	0.23	2.76e-12
Imodium	APC	0.30	6.51e-15
capecitabine	APC	0.30	1.94e-15
Pulmonary function tests	STK11	0.16	2.19e-09

Table 5.1: **Subset of clinico-genetic associations found using the CC setup.** A complete list can be found in Appendix 5.A.

Next, Table 5.3 shows the list of inferred latent features by the H-PFA model. The bias term

Phenotype	**Gene**	β_{qg}	p**-value**
Stage IV Lung Adenocarcinoma	EGFR	0.14	9.90e-14
Pulmonary function tests	STK11	0.11	5.67e-06
Esophagogastroduodenoscopy	ERBB3	0.10	4.62e-06
Adenocarcinoma of lung (disorder)	EGFR	0.09	3.20e-06
Rash and Dermatitis Adverse Event Associated with Chemoradiation	TCF7L2	0.09	1.57e-04
Atrophic	PTEN	0.09	2.06e-05
Esophagogastroduodenoscopy	ALK	0.09	1.43e-05
Stage level 2	ERBB4	0.08	3.34e-05
Positive Surgical Margin	EP300	0.08	1.43e-04
Esophagogastroduodenoscopy	CDH1	0.08	1.33e-04
Esophagogastroduodenoscopy	FLT4	0.08	4.15e-05
Malignant neoplasm of urinary bladder	TERT	0.08	7.91e-05
Potassium Ion	KRAS	0.08	6.72e-06
KRAS gene	KRAS	0.07	9.76e-05

Table 5.2: **Clinico-genetic associations found using the LMM approach.** The associations have been sorted according to the effect size β_{qg} which refers to the linear weight of the regression, as described in Section 5.2.2.

F0 reflects the high rate mutation of the TP53 gene which occur across all types of cancer. The TP53 gene is essential for the production of a protein called tumor protein p53. This protein acts as a tumor suppressor, which means that it regulates cell division by keeping cells from growing and dividing too fast or in an uncontrolled way. Because p53 is essential for regulating cell division and preventing tumor formation, it has been nicknamed the "guardian of the genome" [142]. On top of the bias term F0, H-PFA inferred 19 other latent features. Features F3, F5, and F17 capture complex phenotypes (no somatic mutations involved), whereas F4 and F18 mostly capture somatic mutations. Interestingly, F18 relates Esophagogastroduodenoscopy (a test to examine the lining of the esophagus, stomach, and the beginning of the small intestine) to multiple somatic mutations, which was already revealed by LMM in Table 5.2. The remaining 14 features capture co-ocurrence of somatic mutations and clinical UMLS terms. Some latent features reflect well known relationships in oncology research. To name a few, mutations of gene PIK3CA (captured by F1 and F16) are present in over one-third of breast cancers; such mutations are nowadays known to be oncogenic and also implicated in cervical cancers [74]. Somatic mutations in the triad APC-KRAS-TP53 genes (captured by F0 and F6 together) are prominent in colon cancer [1]. Finally, previous studies have found direct physiological and molecular evidence for a role of gene FOXA1 in controlling cell proliferation in prostate cancer [95], which is accounted for in factor F12.

Figure 5.3 depicts the cancer-specific activation weights π_l^ℓ for each type of cancer ℓ, as described in previous section.The activation of features present strong variations across cancer types. Some features are clearly cancer-specific (F1 and F3 typically activate for breast carcinoma patients; F6, F11 and F15 are typically active for colorectal cancer; F7, F8 and F10 are almost exclusively active for non-small cell lung cancer, etc.), whereas other factors occur in similar proportions across

Feat.	m_k	Phenotypes	Genes
F0.	1946	-	TP53 (0.40)
F1.	460	Simple mastectomy (0.17), Xeloda (0.15), Lumpectomy of breast (0.12), capecitabine (0.09)	PIK3CA (0.31)
F2.	402	Renal function (0.29), Coronary Artery Disease (0.28), Stent, device (0.22), cardiologist (0.21), Urology (0.20), Hydronephrosis (0.16)	MTOR (0.04)
F3.	400	Invasive Ductal Breast Carcinoma (0.57), axillary lymph node dissection (0.38), Simple mastectomy (0.37), Noninfiltrating Intraductal Carcinoma (0.36), Lumpectomy of breast (0.33), Adriamycin (0.29)	-
F4.	392	-	ETV6 (0.22), PTPRD (0.19), ATR (0.19), PTPRT (0.17), BRAF (0.17), ATM (0.17), . . .
F5.	361	Entire intercostal space (0.22), Midclavicular line (0.21), Per Minute (0.20), Prednisone (0.19), Upper Extremity (0.19), Entire head (0.18), Dizziness (0.17), Redness (0.17), Serum (0.17), Bedtime (qualifier value) (0.16), . . .	-
F6.	352	Colorectal (0.39), FOLFOX Regimen (0.29)	APC (0.71), KRAS (0.47)
F7.	350	Lobectomy (0.48), Pulmonary function tests (0.29), Thoracotomy (0.27), Non-Small Cell Lung Carcinoma (0.27)	EGFR (0.09)
F8.	326	Non-Small Cell Lung Carcinoma (0.13), Stage IV Lung Adenocarcinoma (0.10), natural daughter - RoleCode (0.09), pemetrexed (0.08)	KRAS (0.36), STK11 (0.26), KEAP1 (0.20)
F9.	326	Lytic lesion (0.29), Zometa (0.27), Fracture (0.25), Sclerosis (0.25), Bone Lesion (0.23), Bone structure of sacrum (0.23), Hip arthralgia (0.23), Bone structure of ilium (0.19), Palliative Care (0.17)	ATRX (0.04)
F10.	266	Stage IV Lung Adenocarcinoma (0.62), pemetrexed (0.61), Adenocarcinoma of lung (disorder) (0.60), mediastinal lymphadenopathy (0.35)	EGFR (0.30), TP53 (0.18)
F11.	265	FOLFOX Regimen (0.54), KRAS gene (0.45), Folinic Acid-Fluorouracil-Irinotecan Regimen (0.43), Leucovorin (0.43), irinotecan (0.39), Colorectal Carcinoma (0.39), Cold intolerance (0.37), Midclavicular line (0.27), Sigmoid colon (0.27), Colorectal (0.27)	PTPRT (0.05), CARD11 (0.04)
F12.	262	Prostate carcinoma (0.74), adenocarcinoma of the prostate (0.69), Biopsy of prostate (0.62), Extracapsular (0.55), Lupron (0.46), Personal Attribute (0.40)	FOXA1 (0.11), APC (0.06)
F13.	261	Tract (0.52), Malignant neoplasm of urinary bladder (0.51), Gross hematuria (0.44), Incontinence (0.29), Immunotherapy (0.28)	TERT (0.66), KDM6A (0.36)
F14.	169	Lovenox (0.28), Pulmonary Embolism (0.27), Deep Vein Thrombosis (0.23), swollen feet/legs (0.19)	-
F15.	159	Rectum (0.45), Rash and Dermatitis Adverse Event Associated with Chemoradiation (0.28), capecitabine (0.26), Node stage N0 (0.23)	APC (0.14), TCF7L2 (0.14), TSC2 (0.09)
F16.	149	Consistency (0.55), Vagina (0.50), Clinic / Center - Mobile (0.43), Bilateral Salpingectomy with Oophorectomy (0.39), Atrophic (0.39), New medications (0.35), Personal Attribute (0.32), Uterus (0.31), Ovarian (0.30), Bone Mineral Density Test (0.30), Ovary (0.29)	PIK3CA (0.12), PTEN (0.07)
F17.	107	Depression motion (0.95), Structure of long bone (0.82), S3 (sacral segmental innervation) (0.82), pump (device) (0.79), intrahepatic (0.78), Pulse taking (0.74), Midclavicular line (0.73), Entire intercostal space (0.70), Hepatectomy (0.62), Flowcharts (Computer) (0.57)	-
F18.	95	Esophagogastroduodenoscopy (0.12)	POLE (0.65), ROS1 (0.61), DNMT1 (0.59), ATR (0.58), ATM (0.57), FAT1 (0.54), . . .
F19.	41	Optic Nerve (0.90), Gross hematuria (0.90), Dyspepsia (0.57), Lupron (0.45)	AR (0.22)

Table 5.3: **Latent features inferred by H-PFA.** We depict the UMLS terms and genes with highest weights, up until the weight decays more than 50% separately. m_k is the number of patients with each feature active.

cancers, e.g., feature F5 which capture typical adverse effects that manifest for all types of cancer (Prednisone is a synthetic corticosteroid drug which is regularly used to treat certain types of cancer, but has significant adverse effects).

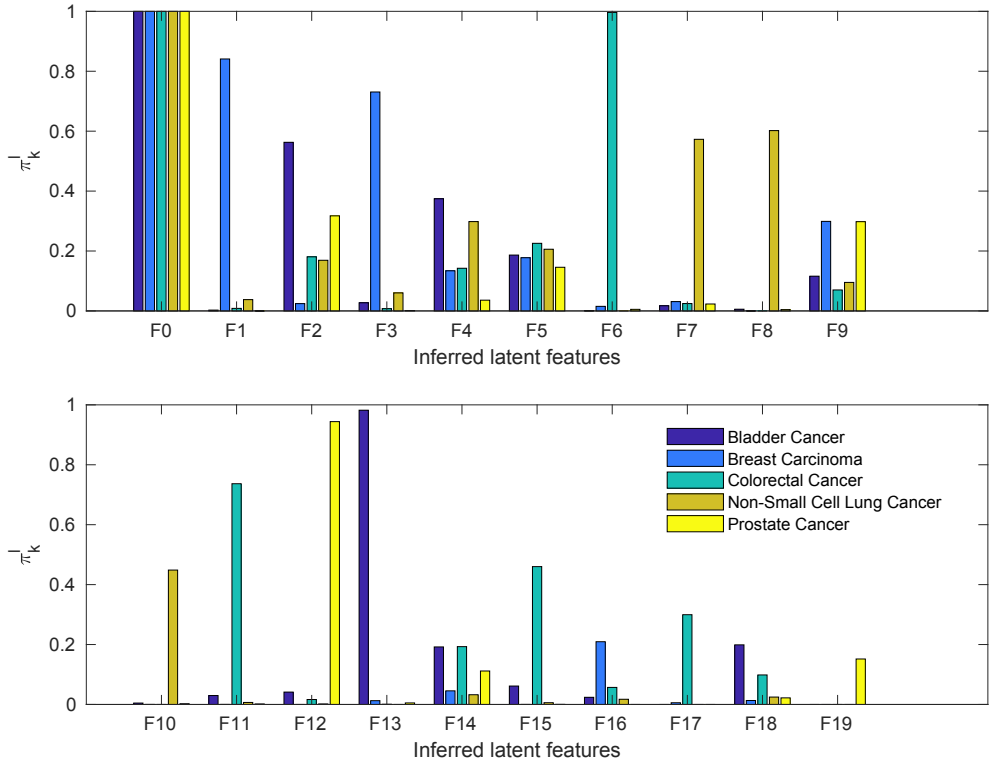


Figure 5.3: **Activation weights π_k^ℓ for each cancer type ℓ inferred by H-PFA.**

Table 5.4 presents a list of the biggest subpopulations found by H-PFA in decreasing size. A subpopulation in this context refers to patients that have the exactly same set of latent features active. Given these subpopulations, we follow the statistical methodology presented in Chapter 4 of running two-sample test comparisons between different sets of subpopulations. In this case, we adopt a one-leave-out comparison strategy, which consists in running two-sample individual tests between each subpopulation against all the rest. We thus obtain different sets of statistically significant components (one set for each comparison), which are listed in Tables 5.5, 5.6, 5.8,5.9, and 5.10.

For each clinical and genetic term, we give both the effect size and significance. Our method is able to provide concise grouping of both clinical terms and somatic mutations. Among the clinical terms, we find both phenotypical terms, as well as names of chemotherapy medications (Adriamycin, Irinotecan, or Leucovorin). Table 5.5 shows cancer-specific clinico-genetic associations. We recover well-known associations (such as APC gene mutation being prominant in colorectal cancer,

Id	Activation pattern	Nr. patients
1.	11010 00000 00000 00000	106
2.	10000 00000 00100 00000	91
3.	11000 00000 00000 00000	69
4.	10000 00000 00000 00000	68
5.	10000 00110 00000 00000	46
6.	10000 01000 01000 00000	40
7.	11010 00001 00000 00000	40
8.	10000 01000 00000 00000	30
9.	10100 00000 00100 00000	30
10.	10000 00000 00010 00000	29
11.	10100 00000 00010 00000	28
12.	10000 00000 10000 00000	27
13.	10000 01000 01000 10000	27
14.	10001 00110 00000 00000	27
15.	10001 00000 00010 00000	26

Table 5.4: **Subpopulations found by H-PFA.**

or STK11 to lung carcinoma), but other associations are more surprising, such as GATA3 gene with bone mineral density.

Finally, H-PFA found several statistically significant sets of associations involving somatic mutations in gene TERT, as shown in Table 5.6. Somatic mutations in the gene promoter of telomerase reverse transcriptase (TERT) have been found in 70–79% of bladder tumors in a multi-institutional study published in European Urology [149]. Table 5.6 shows that TERT mutations are associated to not only malignant neoplasm of urinary bladder (which is not surprising), but also hematuria and hydronephrosis. Hematuria refers to the presence of red blood cells in the urine. Also, hydronephrosis is a condition that typically occurs when the kidney swells due to the failure of normal drainage of urine from the kidney to the bladder. Hydronephrosis is not a primary disease, but results from some other underlying disease (cancer in this case) as the result of a blockage or obstruction in the urinary tract. H-PFA points out to interesting gene relationships (KDM6A, CREBBP, and ARID1A genes) with TERT, which have been partially studied in the literature [141, 167, 97].

5.5 Summary

This chapter proposes a novel BNP procedure for genetic association studies that identifies potentially interesting associations between gene-level mutations and clinical features encoded via UMLS terms from EHRs. We propose a hierarchical Bernoulli process Poisson factor analysis model based on a hierarchical construction of BPs and BePs. The delivered associations are statistically signifi-

cant after correction for multiple hypothesis testing combined with a bootstrapping procedure, which makes the approach particularly robust against false positives. These associations give potentially interesting insights for future research in oncology. Studies like this one can provide us with accurate diagnosis, and ultimately inform us about actionable pathways when considering cancer therapy, where interventions through drug administration can be designed.

Clinical record			**Genetic information**		
Invasive Ductal Breast Carcinoma	0.64	3.54e-49	PIK3CA	0.22	4.38e-07
Simple mastectomy	0.39	4.26e-22	GATA3	0.12	3.08e-05
Noninfiltrating Intraductal Carcinoma	0.34	2.02e-20			
Lumpectomy of breast	0.32	3.94e-17			
axillary lymph node dissection	0.28	3.43e-16			
Fishes	0.22	1.01e-10			
Adriamycin	0.21	4.09e-11			
Bone Mineral Density Test	0.15	2.19e-06			

(a) 100% breast carcinoma

Clinical record			**Genetic information**		
FOLFOX Regimen	0.80	1.52e-29	APC	0.63	3.50e-17
Colorectal	0.37	4.04e-09	TP53	0.42	3.50e-08
Sigmoid colon	0.35	1.88e-09			
Colorectal Carcinoma	0.33	8.49e-08			
Cold intolerance	0.32	2.03e-08			
irinotecan	0.29	8.72e-08			
Leucovorin	0.29	4.90e-07			
Folinic Acid-Fluorouracil-Irinotecan Regimen	0.25	1.19e-05			
Hepatectomy	0.23	3.35e-06			

(b) 100% non-small cell lung cancer

Clinical record			**Genetic information**		
Lobectomy	0.37	4.98e-11	KRAS	0.36	2.73e-08
Non-Small Cell Lung Carcinoma	0.31	5.69e-08	STK11	0.26	1.74e-07

(c) 97% breast carcinoma, 3% non-small cell lung cancer

Clinical record			**Genetic information**		
Lobectomy	0.37	3.60e-07	STK11	0.38	8.62e-08
			KEAP1	0.35	1.17e-07

(d) 100% bladder cancer

Table 5.5: **Clinico-genetic associations found by the H-PFA (1/2).**

Clinical record			Genetic information		
Tract	**0.37**	**3.39e-08**	**TERT**	**0.57**	**4.56e-12**
Malignant neoplasm of urinary bladder	**0.36**	**3.92e-07**	FGFR3	0.50	4.15e-13
Gross hematuria	**0.33**	**2.67e-06**			

(a) 100% colorectal cancer

Clinical record			Genetic information		
Gross hematuria	**0.74**	**1.81e-20**	**TERT**	**0.58**	**5.59e-12**
Malignant neoplasm of urinary bladder	**0.41**	**2.80e-08**	KDM6A	0.31	6.48e-06
Tract	**0.39**	**2.08e-08**			
chain of objects	0.32	1.92e-06			
Hydronephrosis	0.26	2.52e-05			

(b) 100% prostate cancer

Clinical record			Genetic information		
Gross hematuria	**0.39**	**1.72e-07**	**TERT**	**0.64**	**4.86e-13**
Malignant neoplasm of urinary bladder	**0.37**	**9.85e-07**	FGFR3	0.50	1.97e-11
Tract	**0.30**	**2.19e-05**	KDM6A	0.36	1.20e-06
			CREBBP	0.32	3.19e-06

(c) 100% non-small cell lung cancer

Clinical record			Genetic information		
Malignant neoplasm of urinary bladder	**0.56**	**4.24e-12**	**TERT**	**0.67**	**1.79e-14**
Tract	**0.46**	**2.00e-10**	ARID1A	0.45	2.32e-08
Renal function	0.45	5.51e-11			
Gross hematuria	**0.36**	**8.69e-07**			

(d) 100% bladder cancer

Table 5.6: **Clinico-genetic associations found by the H-PFA (2/2).** All these associations involve gene TERT. One same set (depicted in bold) appears in all associations.

5.A Appendix: Complete List of Associations

5.A.1 Case-control setup (CC)

Clinical term	Gene	β_{qg}	p-value
FOLFOX Regimen	APC	0.66	5.12e-113
Leucovorin	APC	0.65	6.61e-65
Hepatectomy	APC	0.65	1.21e-52
irinotecan	APC	0.62	4.39e-47
pump (device)	APC	0.61	2.78e-53
Colorectal Carcinoma	APC	0.61	2.34e-63
Colorectal	APC	0.60	2.31e-73
Folinic Acid-Fluorouracil-Irinotecan Regimen	APC	0.59	1.08e-60
Malignant neoplasm of urinary bladder	TERT	0.55	1.07e-61
S3 (sacral segmental innervation)	APC	0.54	3.58e-35
intrahepatic	APC	0.53	1.03e-30
Tract	TERT	0.50	1.62e-42
Sigmoid colon	APC	0.48	6.82e-38
Gross hematuria	TERT	0.47	1.23e-47
Flowcharts (Computer)	APC	0.47	4.29e-32
Entire intercostal space	APC	0.42	1.33e-41
Unresectable	APC	0.42	1.78e-20
Structure of long bone	APC	0.42	5.86e-29
Midclavicular line	APC	0.41	1.46e-41
KRAS gene	APC	0.39	6.40e-29
Cold intolerance	APC	0.39	1.70e-22
Rectum	APC	0.38	3.88e-26
Data Port	APC	0.37	1.35e-19
pump (device)	TP53	0.35	8.51e-17
Depression motion	APC	0.35	5.62e-22
Avastin	APC	0.35	1.58e-22
KRAS gene	KRAS	0.33	2.44e-23
FOLFOX Regimen	KRAS	0.32	1.96e-32
Stage IV Lung Adenocarcinoma	EGFR	0.32	2.07e-29
Tract	KDM6A	0.31	1.66e-24
Unresectable	TP53	0.31	3.62e-10
capecitabine	APC	0.30	1.94e-15
Imodium	APC	0.30	6.51e-15
Ulcer	APC	0.30	3.96e-12
Flushing	APC	0.30	1.54e-11
Folinic Acid-Fluorouracil-Irinotecan Regimen	KRAS	0.30	1.16e-18
irinotecan	KRAS	0.30	5.60e-14
FOLFOX Regimen	TP53	0.29	4.61e-19
Hepatectomy	TP53	0.29	1.19e-10
S3 (sacral segmental innervation)	TP53	0.29	1.02e-09
Colorectal	KRAS	0.29	2.18e-21

Leucovorin	KRAS	0.29	8.95e-16
Ablation	APC	0.29	5.61e-15
Rash and Dermatitis Adverse Event Associated with Chemoradiation	APC	0.29	5.40e-13
Pulse taking	APC	0.29	2.28e-15
Malignant neoplasm of urinary bladder	KDM6A	0.29	5.14e-27
Potassium Ion	KRAS	0.28	4.17e-14
Renal function	TERT	0.28	8.40e-12
Structure of long bone	TP53	0.27	5.79e-11
Hydronephrosis	TERT	0.27	3.21e-12
Hepatectomy	KRAS	0.27	7.27e-12
Xeloda	APC	0.26	1.53e-11
Midclavicular line	TP53	0.26	6.98e-15
Adenocarcinoma of lung (disorder)	EGFR	0.26	1.49e-22
Leucovorin	TP53	0.26	1.57e-09
Entire intercostal space	TP53	0.26	1.47e-13
Data Port	KRAS	0.26	4.25e-11
Gross hematuria	KDM6A	0.26	4.71e-22
Folinic Acid-Fluorouracil-Irinotecan Regimen	TP53	0.26	8.08e-11
Colorectal Carcinoma	TP53	0.24	1.39e-09
pump (device)	KRAS	0.24	2.59e-11
Sigmoid colon	TP53	0.24	3.38e-09
Potassium Ion	APC	0.24	4.02e-10
Colorectal Carcinoma	KRAS	0.23	3.18e-12
Immunotherapy	TERT	0.23	2.76e-12
intrahepatic	TP53	0.23	1.37e-05
Rectum	TP53	0.22	5.43e-08
Neutrophil count decreased	APC	0.22	8.18e-07
Response process	APC	0.22	9.43e-07
Tract	FGFR3	0.22	1.16e-15
Colorectal	TP53	0.21	5.78e-09
Flowcharts (Computer)	TP53	0.21	1.46e-06
Tract	ARID1A	0.21	5.61e-12
Sigmoid colon	KRAS	0.21	6.84e-10
Flowcharts (Computer)	KRAS	0.21	1.20e-08
Rectal hemorrhage	APC	0.21	1.88e-09
Avastin	TP53	0.21	4.41e-07
Bilateral Salpingectomy with Oophorectomy	PIK3CA	0.20	1.65e-06
Unresectable	KRAS	0.20	6.65e-07
Urology	TERT	0.20	8.42e-09
Depression motion	TP53	0.20	1.43e-06
irinotecan	TP53	0.20	3.11e-05
Combined Modality Therapy	APC	0.20	4.41e-06
hearing impairment	TERT	0.19	1.41e-06
Malignant neoplasm of urinary bladder	FGFR3	0.19	4.27e-15
Gross hematuria	FGFR3	0.18	3.11e-14
Adriamycin	PIK3CA	0.18	4.47e-06
pemetrexed	EGFR	0.18	2.06e-12

Entire intercostal space	KRAS	0.18	2.80e-10
S3 (sacral segmental innervation)	KRAS	0.18	1.25e-05
Non-Small Cell Lung Carcinoma	EGFR	0.18	4.11e-11
Rash and Dermatitis Adverse Event Associated with Chemoradiation	KRAS	0.18	6.24e-06
Creatinine	TERT	0.17	1.24e-05
Bone Mineral Density Test	PIK3CA	0.17	1.65e-06
Hydronephrosis	KDM6A	0.17	1.01e-07
Cold intolerance	KRAS	0.17	2.64e-06
intrahepatic	KRAS	0.17	3.49e-05
Avastin	KRAS	0.17	3.87e-07
KRAS gene	TP53	0.17	1.28e-05
Lobectomy	EGFR	0.16	1.72e-08
Malignant neoplasm of urinary bladder	ARID1A	0.16	3.25e-09
lung lesion	APC	0.16	8.22e-06
Attribution	EGFR	0.16	8.10e-06
Simple mastectomy	PIK3CA	0.16	3.40e-08
Renal function	ARID1A	0.16	8.41e-06
Pleura	EGFR	0.16	4.16e-07
Gross hematuria	ARID1A	0.16	1.34e-08
Pulmonary function tests	STK11	0.16	2.19e-09
Tract	CREBBP	0.16	8.83e-10
Invasive Ductal Breast Carcinoma	PIK3CA	0.15	8.87e-09
Midclavicular line	KRAS	0.15	5.38e-08
capecitabine	KRAS	0.15	1.89e-05
Lumpectomy of breast	PIK3CA	0.15	7.20e-07
Malignant neoplasm of urinary bladder	ERBB2	0.15	1.40e-10
Incontinence	TERT	0.15	1.07e-06
chain of objects	TERT	0.15	1.08e-05
Renal function	KDM6A	0.15	6.54e-06
Stent, device	TERT	0.14	4.23e-06
Superficial	TERT	0.14	1.68e-05
Tibialis anterior muscle structure	EGFR	0.14	1.34e-06
Thoracotomy	STK11	0.14	9.65e-07
Tract	ERBB2	0.14	1.87e-07
Urology	KDM6A	0.13	2.55e-06
Lobectomy	STK11	0.13	1.25e-07
Malignant neoplasm of urinary bladder	RB1	0.13	5.74e-08
Lobectomy	KEAP1	0.13	1.22e-08
Tract	STAG2	0.13	4.54e-07
Malignant neoplasm of urinary bladder	CREBBP	0.12	8.84e-08
Leucovorin	SMAD4	0.12	5.31e-08
Tract	PBRM1	0.12	2.11e-07
Tract	ROS1	0.12	1.89e-05
Bilateral Salpingectomy with Oophorectomy	GATA3	0.12	1.93e-05
Non-Small Cell Lung Carcinoma	STK11	0.12	1.09e-07
Non-Small Cell Lung Carcinoma	KRAS	0.12	8.61e-05
Folinic Acid-Fluorouracil-Irinotecan Regimen	SMAD4	0.11	8.21e-08

Malignant neoplasm of urinary bladder	EP300	0.11	2.57e-07
Pulmonary function tests	KEAP1	0.11	8.29e-07
Colorectal Carcinoma	SMAD4	0.11	4.47e-07
FOLFOX Regimen	SMAD4	0.11	2.03e-10
Tract	FAT1	0.11	3.77e-05
Urology	ERBB2	0.11	1.95e-05
KRAS gene	SMAD4	0.11	4.34e-07
irinotecan	SMAD4	0.11	1.45e-05
Immunotherapy	RBM10	0.11	9.55e-06
Rash and Dermatitis Adverse Event Associated with Chemoradiation	TCF7L2	0.11	2.14e-05
Tract	EP300	0.10	1.24e-05
Superficial	FGFR3	0.10	2.29e-05
Immunotherapy	FGFR3	0.10	4.13e-05
Colorectal	PTPRS	0.10	9.11e-07
Malignant neoplasm of urinary bladder	ATM	0.10	8.03e-05
Non-Small Cell Lung Carcinoma	PTPRD	0.10	9.64e-06
Adenocarcinoma of lung (disorder)	KEAP1	0.10	1.46e-07
Tract	SPEN	0.10	3.32e-05
Gross hematuria	CREBBP	0.10	1.43e-05
Gross hematuria	NSD1	0.10	7.95e-07
Tract	FANCA	0.10	3.20e-05
Gross hematuria	ERBB2	0.10	1.81e-05
Sigmoid colon	TCF7L2	0.10	6.46e-06
Tract	ERBB3	0.10	2.09e-05
Non-Small Cell Lung Carcinoma	KEAP1	0.10	1.02e-06
Invasive Ductal Breast Carcinoma	GATA3	0.10	3.71e-08
Malignant neoplasm of urinary bladder	ERBB3	0.09	6.40e-06
Sigmoid colon	SMAD4	0.09	3.46e-05
Colorectal	ERBB4	0.09	3.22e-05
Adenocarcinoma of lung (disorder)	STK11	0.09	1.91e-05
Simple mastectomy	GATA3	0.09	2.59e-06
pemetrexed	STK11	0.09	1.23e-05
Extracapsular	FOXA1	0.08	3.92e-05
Malignant neoplasm of urinary bladder	PBRM1	0.08	7.37e-05
Malignant neoplasm of urinary bladder	NSD1	0.08	4.75e-05
pemetrexed	KEAP1	0.08	1.31e-05
Colorectal	SMAD4	0.08	1.87e-05
Malignant neoplasm of urinary bladder	BRCA1	0.08	1.80e-04
Colorectal	TCF7L2	0.08	2.29e-05
Stage IV Lung Adenocarcinoma	RBM10	0.08	5.16e-05
FOLFOX Regimen	PTPRS	0.08	1.31e-05
Non-Small Cell Lung Carcinoma	EPHA3	0.08	7.61e-05
Prostate carcinoma	FOXA1	0.07	3.61e-05

Table 5.7: **Complete list of clinico-genetic associations found using the Case-Control Set-up.** β_{qg} refers to the linear weight as described in Section 5.2.2. Associations in bold have also been discovered by the H-PFA.

5.A.2 Hierarchical Poisson Factor Analysis

Clinical record			Genetic information		
Prostate carcinoma	0.63	4.06e-43			
Biopsy of prostate	0.52	1.63e-34			
Extracapsular	0.48	3.54e-32			
adenocarcinoma of the prostate	0.43	3.98e-26			
Personal Attribute	0.33	2.23e-16			
Robotics	0.25	5.86e-13			
Pelvic lymph node group	0.20	2.34e-09			
Lupron	0.20	2.37e-07			
Incontinence	0.17	7.87e-07			
External Beam Radiation Therapy	0.17	1.04e-06			
Positive Surgical Margin	0.14	7.71e-06			

(a) 100% prostate cancer

Clinical record			Genetic information		
Invasive Ductal Breast Carcinoma	0.47	1.86e-11			
Adriamycin	0.37	6.54e-11			
Lytic lesion	0.33	2.45e-09			
Lumpectomy of breast	0.33	4.86e-08			
Zometa	0.32	2.47e-08			
Palliative Care	0.22	8.32e-06			

(b) 32.3% breast carcinoma, 41.2% non-small cell lung cancer, 26.5% prostate cancer

Table 5.8: **Additional clinical associations (complex phenotypes) found by the H-PFA.**

Clinical record			Genetic information		
FOLFOX Regimen	0.77	3.47e-19	APC	0.59	5.34e-11
Rectum	0.59	1.56e-14			
Sigmoid colon	0.43	8.36e-09			
Rash and Dermatitis Adverse Event Associated with Chemoradiation	0.42	1.24e-09			
capecitabine	0.42	5.37e-09			
Colorectal	0.42	1.19e-07			
Folinic Acid-Fluorouracil-Irinotecan Regimen	0.40	1.11e-07			
KRAS gene	0.38	2.12e-07			
Ulcer	0.36	3.64e-08			
irinotecan	0.35	3.10e-07			
Rectal hemorrhage	0.34	3.31e-06			
Avastin	0.32	1.13e-05			
Leucovorin	0.29	2.95e-05			
Combined Modality Therapy	0.26	5.02e-05			
Response to treatment	0.25	4.10e-05			

Table 5.9: **Additional clinico-genetic association found by the H-PFA involving APC gene.** This group of associations was found in a subgroup of 100% bladder cancer patients.

Clinical record			Genetic information		
Stage IV Lung Adenocarcinoma	0.64	3.47e-15	EGFR	0.35	8.65e-06
Adenocarcinoma of lung (disorder)	0.56	3.92e-11			
pemetrexed	0.51	7.89e-10			
Tibialis anterior muscle structure	0.32	3.40e-06			

Table 5.10: **Additional clinico-genetic associations found by the H-PFA involving EGFR gene.** This group of associations was found in a subgroup of 100% non-small cell lung cancer patients.

6

Flexible Indian Buffet Process Priors for Understanding International Trade

This chapter presents Bayesian nonparametric (BNP) latent factor models specially suitable for exploratory analysis of high-dimensional count data. First, we perform a non-negative doubly sparse matrix factorization that has two main advantages: on the one hand, we are able to better approximate the row marginal input distribution of the observation matrix; on the other hand, the inferred topics are also sparse, and thus, easier to interpret. By combining the stable-Beta process [201] with the restricted Indian buffet process [43], we increase the model flexibility, allowing for a full spectrum of sparse solutions in the latent space. Second, we propose a dynamic Poisson factorization model based on the Markov Indian buffet process [56] with varying activation of the latent features over time. We demonstrate the usefulness of our approaches in the analysis of countries' economic structure based on two different databases of export portfolios, ranging from 1964 to 2010. Compared to other approaches, empirical results show our model's ability to give easy-to-interpret information and better capture the sparsity structure underlying data.

6.1 Introduction

In this chapter, we focus on finding an interpretable representation for international trade, a topic of central interest to the theory of economic growth and, in particular, to the recently introduced concept of *economic complexity* [88, 83, 192]. Our objective is to help explain the productive structure and competitiveness of world economies based on the thresholded revealed comparative advantage (RCA) matrix [16]. This is a well-established normalized metric correcting for size of economies, defined as:

$$\text{RCA}_{nd} = \frac{E_{nd} / \sum_p E_{nd}}{\sum_n E_{nd} / \sum_{n,d} E_{nd}}, \tag{6.1}$$

$$x_{nd} = \begin{cases} 1, & \text{if RCA}_{nd} \geq 1 \\ 0, & \text{otherwise} \end{cases}. \tag{6.2}$$

where E_{nd} is the raw export of product d by country n in dollars. $\text{RCA}_{cp} > 1$ indicates that country c's share of product p is larger than the product's share of the entire world market, thus "revealing" a comparative advantage of the country in the corresponding product. Let $\mathbf{X}$ be the resulting $N \times D$ sparse high-dimensional matrix reflecting the relative export advantages of N countries at exporting each of the D products.

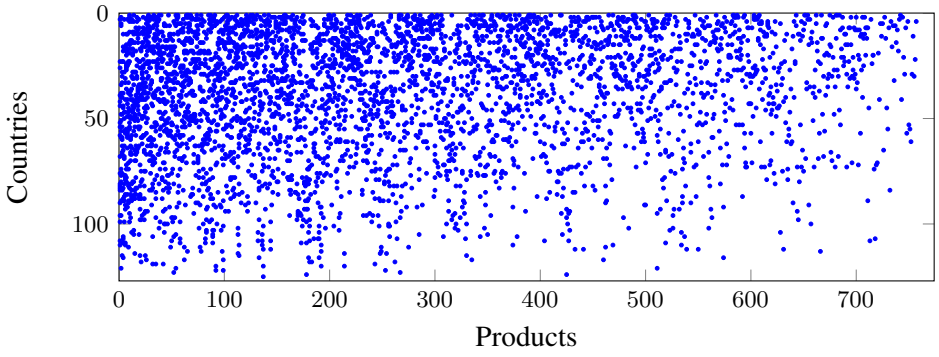


Figure 6.1: **Country-product matrix constructed from real trade data in 2010.** A non-zero entry reflects a relative advantage of a country at exporting a given product. Columns and rows have been arranged in decreasing order according to the number of ones per row and column.

An interesting property of such data is its approximately triangular sparse structure after re-ordering of rows and columns, as shown in Figure 6.1. Rather than a niche economic paradigm where countries focus in the production of a few specialized products (this would result in a block-diagonal matrix), data suggests that countries have different diversity degrees in their export portfolios, and thus different trade strategies and skills. Our objective herein is twofolds: first, we want to find an underlying representation that is easy to understand and able to capture this triangular structure in

the input data for a single year; second, we want to study the evolution of such matrix over time.

This chapter proposes Poisson sparse non-negative matrix factorization models which place Indian buffet process (IBP)-based priors over the country-feature matrix to learn the appropriate number of hidden features from the data to deal with a static and dynamic scenario. To enforce sparsity in the feature-product matrix, we use a component-wise gamma prior with a shape parameter smaller than one. This prior has a similar effect to the spike and slab prior, but is more amenable for inference.

In the static scenario (single year), we additionally modify the IBP prior according to two recent advances [201, 43]. First, we use the tree-parameter Indian buffet process (3P-IBP) formulation from [201] that allows for a different degree of sharing between features, and an eventual power-law distribution over the feature weights. Such flexible prior allows to capture different kinds of realities, from a world in which countries with few skills focus on different types of products, to a world in which poor countries have a strong overlap in export portfolios. Second, we also rely on the restricted Indian buffet process (R-IBP) formulation from [43], which allows for a general marginal distribution over the number of active features per row. Indeed, a priori we expect countries to exhibit different degrees of diversification in the exports, which translates in different amounts of active latent features per row. Poor countries might have very little features active, whereas other diversified countries might have almost all features active.

In the dynamic scenario (multiple years), we extend the basic Bernoulli process Poisson factor analysis (BeP-PFA) model in order to analyze the evolution of countries' capabilities (latent features) over time. Based on the Markov Indian buffet process (mIBP), we design a BNP model in which countries are allowed to acquire or loose capabilities along the years according to a Markovian structure. This temporal perspective is particularly useful in terms of policy recommendation or future export predictions, and allows us to build a finite state machine of latent binary patterns.

The inference for both the static and dynamic models, which we denote as Three-parameter restricted Bernoulli process Poisson factor analysis (3RBeP-PFA) model and Dynamic Bernoulli process Poisson factor analysis (dBeP-PFA) model respectively, is based on Markov chain Monte Carlo (MCMC) approaches. We use auxiliary variables to make our models conditionally conjugate, together with Metropolis-Hasting within Gibbs and Adaptive Rejection Metropolis sampling [126] in the static case, and forward-filtering backward-sampling (FFBS) [56] in the time-varying case.

In the experimental section we use the proposed models to analyze multiple international trade databases. Our models exhibit similar predictive accuracy than other approaches, while it outperforms them in terms of interpretability strength. In the static scenario, the proposed model is the best one at capturing the input triangular structure of the data. Both a quantitative and qualitative analysis

are provided for export data from 2010. We further demonstrate the usefulness of our models to analyze the temporal dynamics of countries' exports between 1964 and 2010.

6.2 Flexible IBP Extensions

6.2.1 Three-Parameter Indian Buffet Process

In the 3P-IBP, the feature weights follow a more flexible distribution that covers power-law behaviors [201]. This can be achieved by replacing the beta process directing measure in the IBP by a stable-beta process:

$$\mu \sim \text{stable-BP}(1, \alpha, H), \qquad \mathbf{Z}_{n\bullet} \sim \text{BeP}(\mu, f) \tag{6.3}$$

As its name indicates, this process can be fully specified by three parameters: α is the same mass parameter from the IBP that controls the *a priori* expected total number of non-zero entries in matrix $\mathbf{Z}$. Additionally, the stability exponent $\sigma \in [0, 1)$ controls the power-law behavior of the model (weight decay), and $c > -\sigma$ is the concentration parameter that affects the *a priori* number of ones per column (sharing degree per column). When $c = 1$ and $\sigma = 0$, we recover the standard IBP.

Using the usual culinary metaphor of customers entering an Indian buffet restaurant and sequentially choosing dishes from an infinite buffet, the 3P-IBP generalizes as follows:

- Customer 1 tries Poisson(α) number of dishes.
- Customer $n + 1$ tries:
  - each dish with probability $\frac{m_k - \sigma}{n+c}$ for each one that has been previously tried; m_k is the number of customers who previously sampled from dish k.
  - Poisson $\left(\alpha \frac{\Gamma(1+c)\Gamma(n+c+\sigma)}{\Gamma(n+1+c)\Gamma(c+\sigma)}\right)$ new dishes.

In such process, the number of hidden features is expected to grow as $\mathbb{O}(N^\sigma)$ where N is the number of samples. By introducing parameters c and σ, $\mathbf{Z}$ can have a more flexible structure, regardless of the sparsity density which is controlled by α. Compared to the IBP, the 3P-IBP gives more flexibility on the feature weights, but has the disadvantage that the number of ones per-row is still a priori Poisson distributed, which might not always be desirable, particularly in our analysis of international trade.

6.2.2 Restricted Indian Buffet Process

The R-IBP is a recently developed model that allows an arbitrary prior distribution to be placed over the number of active features underlying each observation [43]. A natural way to build such process

is to replace the underlying Bernoulli processes in the IBP by *restricted* Bernoulli processes defined as:

$$\text{R-BeP}(\mathbf{Z}_{n\bullet}; \mu, f) = f\left(J_n\right) \cdot \frac{\prod_{k=1}^{\infty} \pi_k^{z_{nk}} (1-\pi_k^{1-z_{nk}}) \mathbb{1}(\sum_K z_{nk} = J_n)}{\sum_{\boldsymbol{z}' \in \mathcal{Z}} \prod_k \pi_k^{z_k'} (1-\pi_k)^{(1-z_k')} \mathbb{1}(\sum_K z_k' = J_n)} \tag{6.4}$$

where the directing measure $\mu = \sum_k \pi_k \delta_{\theta_k}$, π_k and θ_k are the feature weight and location corresponding to each latent feature k, α is the same mass parameter of the IBP, J_n is the number of ones per row n, $\mathcal{Z}$ is the set of all possible binary vectors, and f is the a priori distribution over the number of active features per row. The R-IBP can thus be formulated as:

$$\mu \sim \text{BP}(1, \alpha, H), \qquad \mathbf{Z}_{n\bullet} \sim \text{R-BeP}(\mu, f) \tag{6.5}$$

We then have two degrees of freedom α and f to control for sparsity degree and sparsity structure of $\mathbf{Z}$. Note that columns are not exchangeable anymore, i.e., the parameter f creates correlation among the features, which has to be handled during inference.

The intuition behind the R-IBP can be explained using the previously stated culinary metaphor. Customers in the R-IBP have varying degrees of hunger: some of them sample from many dishes in the buffet, while others only taste a reduced set of dishes. This interpretation makes specially sense in our international trade application, where developed countries are known to have more assets, and thus are expected to exhibit a higher number of latent features (capabilities) compared to poor countries.

6.3 Static Scenario

6.3.1 Three-Parameter Restricted Bernoulli Process Poisson Factor Analysis

To decouple sparsity density and sparsity structure in the latent matrix, we combine the advantages of both the R-IBP and 3P-IBP into a single prior,

$$\mathbf{Z} \sim \text{3R-IBP}(\alpha, c, \sigma, f) \tag{6.6}$$

where the mass parameter α controls the sparsity degree of matrix $\mathbf{Z}$, c is the concentration parameter that accounts for the degree of sharing between features, σ is the stability exponent responsible for the power-law behavior of the feature weights, and f is the a priori distribution over the number of ones per row. By doing so, we expect our model to be able to find highly-specific and easy-to-

interpret hidden features. In our application, we should be able to find "high-tech" features which involve only a few products and are active for a small number of countries, which would be consistent with the economic literature [82].

Let $\mathbf{X} \in \mathbb{N}^{N \times D}$ be our input matrix of N samples and D dimensions. Using the three-parameter restricted IBP prior, we build an infinite latent feature model for count data with Poisson likelihood and Gamma-distributed factors as follows

$$x_{nd} \sim \text{Poisson}\big(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}\big), \tag{6.7}$$

$$B_{kd} \sim \text{Gamma}\Big(\alpha_B, \frac{\mu_B}{\alpha_B}\Big), \tag{6.8}$$

where α_B and μ_B are the shape and mean parameters of the prior Gamma distribution for each element of matrix $\mathbf{B}$. In this model, both matrices $\mathbf{Z}$ and $\mathbf{B}$ are non-negative and sparse, which makes the inferred latent variables very easy to interpret. In particular, sparsity in matrix $\mathbf{B}$ can be induced simply by choosing $\alpha_B \ll 1$.

In our particular application of international trade, we have N countries, D products and K^+ non-empty latent features to be inferred. A given row $\mathbf{Z}_{n\bullet}$ captures which latent features are active for country n. $\mathbf{B}$ represents the effect of each latent feature on every product. For instance, if a latent feature k is active for a certain country, all products having high values in vector $\mathbf{B}_{k\bullet}$ will be more likely to be exported by that country[1].

We further restrict the model in two ways. First, we use a bias term, labeled as feature $F0$ that is active for all the countries. This bias term is not sparse, and will capture the general ad-hoc trading that might exist in a country punctually but not as a trend, and thus does not constitute a feature. A bias term allows for the active features to be sparser and more interpretable, as well as reducing the number of features that are necessary to explain the whole database. Such approach has already been followed in [182, 159] to alleviate identifiability problems in the inferred solution. Second, we rely on a negative binomial distribution for f in (6.6). The negative binomial distribution is best understood as an overdispersed Poisson [226]. Hence it will naturally allow for countries to exhibit a much variable range of active features, from not having any additional feature on top of the bias, to having all the latent features active. These two extremes are less likely to happen in the standard IBP model.

We call the whole model three-parameter restricted Bernoulli process Poisson factor analysis (3RBeP-PFA). This model can be seen as a probabilistic extension of non-negative matrix factor-

[1]The vector $\mathbf{B}_{k\bullet}$ corresponds to the k-th row of matrix $\mathbf{B}$.

ization where the number of latent features is not fixed a priori, both matrices are sparse, and soft-constraints on the latent sparsity structure are imposed through the prior.

6.3.2 Inference

Since exact computation of the posterior distribution for the latent variables is intractable, we resort to a Markov Chain Monte Carlo (MCMC) approach. In particular, our algorithm uses Gibbs sampling together with Metropolis-Hasting (MH) and Adaptive Rejection Metropolis Sampling (ARMS) [126]. We use a finite-dimensional approximation for the latent measure $\boldsymbol{\pi}$ by allowing at most K features.

For each observation x_{nd}, we introduce the auxiliary variables $x'_{nd,1}, \ldots, x'_{nd,K}$ such that $x_{nd} = \sum_{k=1}^{K} x'_{nd,k}$, and $x'_{nd,k} \sim \text{Poisson}(Z_{nk}B_{kd})$ for $k = 1, \ldots, K$. Given such auxiliary variables, the model is conditionally conjugate, and a Gibbs sampler can be derived straightforwardly. The complete sampling algorithm is described in Algorithm 5. Further details regarding the slice sampler for the BeP-PFA based on the one-parameter IBP can be found in Section 8.

Algorithm 5 A single iteration of the MCMC inference procedure for the 3RBeP-PFA model.

1: Sample each element of matrix $\mathbf{Z}$ using inclusion probabilities [5, 43].
2: Sample latent measure $\boldsymbol{\pi}$ using MH steps [43]. The ARMS is needed to sample the feature weights of completely new features.
3: Sample each element of $\mathbf{B}$ and $\mathbf{X}'$ from their conditional distributions.
4: Sample hyperparameter α according to [50].

In the case of the 3RBeP-PFA model, the inference algorithm proceeds as follows:

1. Sample $\mathbf{Z}$ element-wise from each conditional probability distribution, according to [43]. Inclusion probabilities are combined with f and the data likelihood $p(\mathbf{X}|\mathbf{Z}, \mathbf{B})$.

2. Sample the latent measure $\boldsymbol{\pi}$. Since the beta process is not conjugate to the restricted Bernoulli process, we cannot directly Gibbs sample $\boldsymbol{\pi}$. Instead, we use a Metropolis-Hastings within Gibbs sampling step. We use the posterior distribution from the standard IBP as a proposal distribution $\mathcal{Q}$. The posterior distribution of weights in the case of non-empty features is given by

$$\mathcal{Q}(\pi_k|\mathbf{Z}_{\bullet k}) \propto \text{Beta}(\frac{\alpha c}{K} + \sum_{n=1}^{N} z_{nk} - \sigma,\ N - \sum_{n=1}^{N} z_{nk} + c + \sigma) \tag{6.9}$$

   In the case of new (empty) features, the posterior distribution is not conjugate anymore, so we need to use additional sampling techniques such as Adaptive Rejection Metropolis Sam-

pling [126]. Note that the posterior of the latent measure is not a stable-beta process anymore. The posterior distribution from which we should sample is given by

$$\mathcal{Q}(\pi_{k_{new}}|\mathbf{Z}_{\bullet k}) \propto \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \pi_{k_{new}}^{-\sigma-1}(1-\pi_{k_{new}})^{c+N+\sigma-1} \tag{6.10}$$

The acceptance probability is given by

$$a = \frac{p(\boldsymbol{\pi}', \mathbf{Z})}{p(\boldsymbol{\pi}, \mathbf{Z})} \frac{\mathcal{Q}(\boldsymbol{\pi}|\mathbf{Z})}{\mathcal{Q}(\boldsymbol{\pi}'|\mathbf{Z})} \tag{6.11}$$

$$= \frac{\prod^N \text{R-BeP}(\mathbf{Z}_{n\cdot}; \boldsymbol{\pi}', f)}{\prod^N \text{R-BeP}(\mathbf{Z}_{n\bullet}; \boldsymbol{\pi}, f)} \frac{p(\boldsymbol{\pi}')}{p(\boldsymbol{\pi})} \frac{\mathcal{Q}(\boldsymbol{\pi}|\mathbf{Z})}{\mathcal{Q}(\boldsymbol{\pi}'|\mathbf{Z})} \tag{6.12}$$

where

$$\text{R-BeP}(\mathbf{Z}_{n\bullet}; \boldsymbol{\pi}, f) = f(J_n) \frac{\prod_{k=1}^{\infty} \pi_k^{z_{nk}}(1-\pi_k^{1-z_{nk}}) \mathbb{1}(\sum_K z_{nk} = J_n)}{\sum_{z' \in \mathcal{Z}} \prod_k \pi_k^{z_k'}(1-\pi_k)^{(1-z_k')} \mathbb{1}(\sum_K z_k' = J_n)} \tag{6.13}$$

Let us call $D_{J_n}^K$ the denominator in (6.13). This value can be computed easily using a dynamic programming approach. Indeed, we can exploit the recursive form $D_{J_n}^K = (1-\pi_K) D_{J_n}^{K-1} + \pi_K D_{J_n-1}^{K-1}$.

3. Gibbs sample $\mathbf{B}$, the auxiliary variables $\mathbf{X}'$, and hyperparameter α in the same way as in the standard IBP model.

6.4 Time-varying Scenario

6.4.1 Dynamic Bernoulli Process Poisson Factor Analysis

In the following, we focus on the temporal dynamics of the hidden features. Our objective is to monitor which capabilities each country acquires or loses over time, in order to further understand the development paths of an economy. We extend the basic BeP-PFA based on the infinite factorial hidden Markov model [56]. Note that we take the BeP-PFA as a starting point instead of the improved 3RBeP-PFA model for simplicity reasons in the modeling and for inference speed; an extension based on the 3RBeP-PFA is in the agenda for future work. Let $\mathbf{X}^{(t)}$ be our country-product matrix at timestamp $t = 1, \ldots, T$. The new likelihood function for each element in this matrix can be written as:

$$x_{nd}^{(t)} \sim \text{Poisson}\left(\mathbf{Z}_{n\bullet}^{(t)} \mathbf{B}_{\bullet d}\right) \tag{6.14}$$

where $\mathbf{Z} \in \{0,1\}^{N \times K \times T}$ is a three-dimensional matrix for N countries, $K \to \infty$ inferred latent features, and T timestamps. Our model assumes that each latent feature k has independent and constant Markov dynamics over time, shared for all countries. Let Q_k be the transition matrix for feature k, defined as:

$$Q_k = \begin{pmatrix} 1 - a_k & a_k \\ 1 - b_k & b_k \end{pmatrix} \tag{6.15}$$

where a_k is the activation probability of feature k, and b_k is the probability of staying active, sometimes called the persistence parameter.

The feature activation matrix $\mathbf{Z}$ is generated via a collection of mIBPs with shared transition matrices and hyperparameters across countries. In particular, for each country n, we have $\mathbf{Z}_{n\bullet}^{(\bullet)} \sim$ mIBP(α), which is equivalent to the following generative process when $K \to \infty$:

$$\begin{aligned} a_k &\sim \text{Beta}\Big(\frac{\alpha}{K}, 1\Big), && (6.16)\\ b_k &\sim \text{Beta}(\gamma, \delta), && (6.17)\\ z_{nk}^{(t)} | a_k, b_k &\sim \text{Bernoulli}\left(a_k^{1-z_{nk}^{(t-1)}} b_k^{z_{nk}^{(t-1)}}\right) && (6.18)\\ z_{nk}^{(0)} &= 1 && (6.19) \end{aligned}$$

The prior for each element $\mathbf{B}_{kd}$ is the same as in the BeP-PFA model, i.e., a Gamma distribution with shape parameter smaller than one in order to enforce sparsity in the latent features. We call the complete model dynamic Bernoulli process Poisson factor analysis (dBeP-PFA).

Since it is not possible to directly extent the 3RBeP-PFA to the dynamic scenario, we resort to a one-parameter IBP prior,

6.4.2 Inference

To make inference tractable, we make the same assumption as in the static scenario, namely, that each observation $x_{nd}^{(t)}$ is equal to the sum of K Poisson-distributed auxiliary random variables, i.e., $x_{nd}^{(t)} = \sum_{k=1}^{K} r_{nd,k}^{(t)}$ where $r_{nd,k}^{(t)} \sim \text{Poisson}(z_{nk}^{(t)} B_{kd})$ for $k = 1, \ldots, K$. We adopt the convention that a zero-rate Poisson is equivalent to a delta at zero[2]. Notice that vector $\mathbf{Z}_{n\bullet}^{(\boldsymbol{t})}$ works as a *mask* on the multinomial auxiliary variables, and therefore there exists a very strong correlation between $\mathbf{Z}_{n\bullet}^{(\boldsymbol{t})}$ and $\boldsymbol{r}_{\boldsymbol{n\bullet,\bullet}}^{(\boldsymbol{t})}$ for each country n and timestep t. For interpretability purposes, we force the first feature to be active for all samples. Such feature works as a bias term, it captures the a priori count rates (average

[2]This corresponds to the assumption that $0 \cdot \log(0) = 0$.

values) for each dimension d, and avoids numerical problems.

As explained in [187], a naïve Gibbs sampling in a time series model is typically very slow due to potentially strong coupling between successive time steps. Thus, we resort to a blocked Gibbs sampler which fixes all but one column of $\mathbf{Z}$, and runs a forward-filtering backward-sampling sweep on the remaining column [56]. Matrix $\mathbf{B}$ can be sampled exactly conditioned on all other variables.

We start with an initial $\mathbf{Z}$ matrix and sample a_k, b_k $\forall k$. Next, conditional on our initial $\mathbf{Z}$ and our observations $\mathbf{X}$, we sample the feature effect matrix $\mathbf{B}$. We then start an iterative sampling scheme described in Algorithm 6. Further details concerning the complete conditionals and the FFBS procedure can be found in the Appendix 8.

Algorithm 6 Slice sampler for the dBeP-PFA

1: Sample the auxiliary slice variable μ. This might involve extending the representation of the other latent variables, e.g., $\boldsymbol{r}$ and $\mathbf{B}$.
2: For all the represented features, sample $\mathbf{Z}$, $\boldsymbol{r}$ and $\mathbf{B}$,
3: Sample the model hyperparameters α, γ, δ,
4: Remove all unused features.

6.5 Results

Data description. To illustrate the usefulness of our model, we consider two publicly available trade datasets, both involving N countries and D products. The first considered database is the UN COMTRADE Standard international trade classification (SITC) rev.2 dataset, which disaggregates products to the four digit level, provided by the team of the Observatory of Economic Complexity. In order to clean unreliable or inadequately classified data, we restrict the dataset to the same countries that were used in the Atlas of Economic Complexity [83]. This leaves us with data on 126 countries and 744 products.

The second database that we use is the Harmonized system (HS) rev. 1992 classification disaggregated to six digit level (4890 products). The original data was collected by UN COMTRADE, and was further cleaned by the team of the Observatory of Economic Complexity (the HS data were also cleaned by the BACI team). Both databases are available at `http://atlas.media.mit.edu/en/resources/data/`.

The biggest difference between both datasets is the number of considered products and sparsity density, as stated in Table 6.1. Note that $D \gg N$, which is not the common scenario for latent feature models. The data matrix is binary and represents the RCA of countries, which is a normalized common measure in economics [16]. Basically, an entry x_{nd} in the country-product matrix $\mathbf{X}$ equals

	N	D	Nr. entries (2010)	Nr. entries (all years)	Sparsity
SITC	126	744	16k	1.1M	0.17%
HS	123	4890	77k	1.5M	0.13%

Table 6.1: **Databases on International Trade considered in this chapter.** The available time windows for both databases are (1964-2014) for the SITC data and (1995-2014) for the HS database.

one when country n has a relative advantage at exporting product d, and zero otherwise. Even if the matrix is binary, we may use a Poisson likelihood because of the high degree of sparseness in data. Such approximation has already been adopted successfully in the case of recommendation systems [67].

Experimental setup. Regarding the models' hyperparameters, we choose a Gamma prior over the concentration parameter α with shape and scale parameters equal to one. The concentration parameter α is sampled as described in Section 6.3.2. We set α_B to 0.01 to induce sparsity, and μ_B equal to one. In the static setting, we choose a flexible marginal prior f as Negative-Binomial(r, p), with $r = [1, 2]$, and $p = [0.1, 0.3, 0.5]$. The results are equivalent using any of these priors. Specifically, we report the results for $r = 1$ and $p = 0.1$. Additionally, we run experiments for each combination of $c = [1, 10, 20, 50]$ and $\sigma = [0, 0.25, 0.5, 0.75, 1]$. Even if the results did not vary considerably when changing those hyperparameters, setting $c > 1$ and $\sigma > 0$ allows for a higher *a priori* sparseness in the latent features and potential power-law behaviors in the feature weights respectively. All figures and tables in the static case correspond to $c = 50$, and $\sigma = 1$. Finally, for the dynamic scenario, we assume a Beta prior for b_k, with both γ and δ equal to one (which is equivalent to the uniform distribution).

6.5.1 Static Scenario

Quantitative Evaluation

We first perform a quantitative evaluation of our model in terms of predictive accuracy, interpretability strength and ability to capture the row marginal distribution of the input, using data from 2010. Simulations are run for 10 different train-test splits with a proportion of 90-10% entries. The burn-in period for the MCMC inference algorithm is 30,000 iterations, and results are averaged using the last 1,000 posterior samples. Table 6.2 compares our model against probabilistic matrix factorization (PMF) [134], non-negative matrix factorization (NNMF) [185], the standard BeP-PFA, and the sparse sparse Bernoulli process Poisson factor analysis (sBeP-PFA) which uses $\alpha_B \ll 1$.

Metric	PMF	NNMF	BeP-PFA	SBeP-PFA	3RBeP-PFA
Log Perplexity	1.68 ± 0.01	1.61 ± 0.01	$\mathbf{1.59 \pm 0.04}$	3.26 ± 0.17	1.62 ± 0.01
Coherence	-264.60 ± 4.74	-263.27 ± 7.45	-149.36 ± 7.56	-178.44 ± 4.50	$\mathbf{-140.51 \pm 2.73}$

(a) 2010 SITC database ($N = 126$, $D = 744$)

Metric	PMF	NNMF	BeP-PFA	SBeP-PFA	3RBeP-PFA
Log Perplexity	1.48 ± 0.01	$\mathbf{1.47 \pm 0.01}$	1.58 ± 0.01	2.56 ± 0.12	1.57 ± 0.02
Coherence	-264.73 ± 3.11	-264.67 ± 6.22	-148.91 ± 10.57	-168.39 ± 13.16	$\mathbf{-134.51 \pm 4.43}$

(b) 2010 HS database ($N = 123$, $D = 4890$)

Table 6.2: **Quantitative Evaluation of Accuracy and Interpretability for 3RBeP-PFA.** We compare PMF, NNMF, BeP-PFA, sBeP-PFA and 3RBeP-PFA in terms of mean and standard error of the test log-perplexities, and topic coherence.

Accuracy. We use perplexity to measure predictive accuracy, i.e. the harmonic mean of the inverse test log likelihood (the lower, the better), defined as:

$$\text{Perplexity}(\mathbf{X}_{\text{test}}) = \exp\left(-\frac{\ln p(\mathbf{X}_{\text{test}}|\mathbf{X}_{\text{train}})}{N_{\text{test}}}\right) \tag{6.20}$$

$$\ln p(\mathbf{X}_{\text{test}}|\mathbf{X}_{\text{train}}) \approx \text{Poisson}\Big(\mathbb{E}\left[p(\mathbf{Z}|\mathbf{X}_{\text{train}}\right]\ \mathbb{E}\left[p(\mathbf{B}|\mathbf{X}_{\text{train}})\right]\Big) \tag{6.21}$$

All models present similar perplexity, except the sBeP-PFA model, in which the sparseness restriction degrades its performance significantly. Our 3RBeP-PFA has the same sparse restriction, but it has a more flexible prior that it is able to compensate the penalty in perplexity and perform close to the non-sparse models, i.e. PMF, NNMF and BeP-PFA. The combination of the negative binomial and the stable-Beta process allows to match the perplexity performance of non sparse methods, but keeping the results interpretable, as we illustrate in the next paragraphs.

Interpretability. In order to assess semantic quality, we rely on the coherence [51], which is an often-used metric in topic modeling literature. The coherence C_k of a feature is defined as:

$$C_k = \sum_{m=2}^{M}\sum_{l=1}^{m-1} \log \frac{R(v_m^k, v_l^k) + 1}{R(v_l^k)} \tag{6.22}$$

where v_i^k is the i-th product with highest weight in factor k, and M represents how many top products should be evaluated, here we take $M = 20$ products. Also $R(x)$ refers to the number of countries exporting product x, and $R(x, y)$ is the number of countries exporting both products x and y. The closer coherence is to zero, the better. The 3RBeP-PFA outperforms the BeP-PFA and sBeP-PFA by far, making it specially suitable for data exploration in high-dimensional count scenarios. The

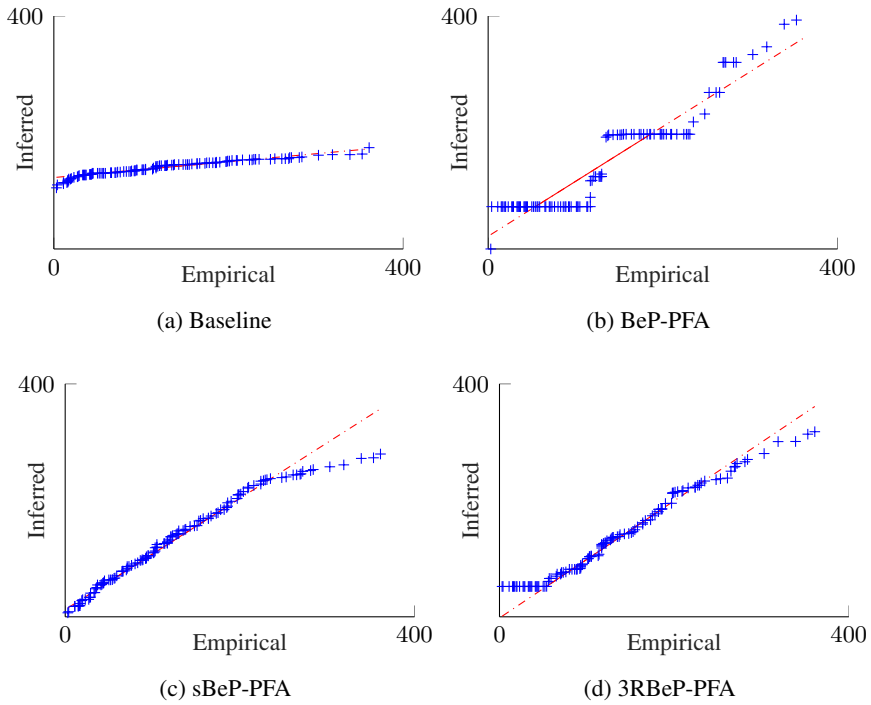


Figure 6.2: **Capturing Sparsity Structure.** Comparison of the empirical and inferred distribution of the number of ones per row in the input matrix, using qq-plots.

non-sparse methods present low coherence, as expected.

Sparsity structure. Next, we evaluate the ability of our model to fit the input distribution of countries' diversity, i.e., the number of ones per-row in $\mathbf{Z}$. Figure 6.2 presents a comparison of our 3RBeP-PFA model with the standard BeP-PFA, the sBeP-PFA model, and a simple binomial model described in [82]. The latest one assumes binary matrices $\mathbf{Z}$ and $\mathbf{B}$, a finite number of capabilities K, and uniform activation probabilities p_{cc} and p_{cp} for all country-capability and capability-product combinations[3]. We measure the "proximity" of the empirical and predicted distribution of the number of ones per row for all models through a qq-plot. The sBeP-PFA underfits the distribution for higher values, e.g., it predicts a lower number of countries with high number of exports, in contrast to the 3RBeP-PFA model[4].

[3]Parameters in this baseline model are chosen to best fit the diversity input histogram.
[4]The horizontal line for the 3RBeP-PFA corresponds to countries only having the bias term $F0$.

Qualitative Evaluation

Interpretability Table 6.3 compares two similar latent features (capabilities) inferred by our 3RBeP-PFA model and using a standard singular value decomposition (SVD). In each case, we report a sorted list of products and their corresponding factor weights. We only include those products whose weight is large enough (at most 30% lower than the maximum weight). Compared to SVD, our model is able to give much shorter and concise descriptions, as weights decrease at a faster pace and the largest weights are larger. Moreover, products in the SVD list come from a mixture of farming and technological elements, whereas the 3RBeP-PFA list is more homogeneous. We conclude that our approach enhances interpretability of the latent factors in terms of both conciseness and precision.

Top Products with weights $> 30\%$	B_{kd}
Bovine	0.49
Misc. Refrigeration Equipment	0.43
Radioactive Chemicals	0.41
Blocks of Iron and Steel	0.41
Rape Seeds	0.40
Animal meat, misc	0.39
Refined Sugars	0.38
Misc. Tire Parts	0.38
Leather Accessories	0.38
Liquor	0.38
Bovine meat	0.38
Embroidery	0.37
Unmilled Barley	0.37
Dried Vegetables	0.36
Textile Fabrics Clothing Accessories	0.36
Horse Meat	0.35
Iron Bars and Rods	0.35
Analog Navigation Devices	0.35

(a) SVD

Top Products with weights $> 30\%$	B_{kd}
Misc. Animal Oils	0.78
Bovine and Equine Entrails	0.72
Bovine meat	0.68
Preserved Milk	0.63
Equine	0.62
Butter	0.58
Misc. Animal Origin Materials	0.57
Glues	0.56

(b) 3RBeP-PFA

Table 6.3: **Qualitative comparison for a similar latent feature for SVD and 3RBeP-PFA.**

Table 6.4 presents a complete list of all capabilities inferred by our model. For each capability k, we report the averaged number of countries $\bar{m_k}$ that have it, the top-5 products with highest weights B_{kd} and a *representative country*, which we define as the country that has the least number of capabilities among those that possess capability k. We also report the average number of active capabilities $\bar{J}_n$ for each representative country n. Clearly, our model is highly interpretable in the sense that it is able to differentiate capabilities according to the elements required in the production of products. For instance, we can clearly associate capability F2 to the presence of animals and farming, or capability F7 to the elements required in the production of automotive parts. Furthermore, capabilities for which the corresponding representative countries have a high number of active capa-

Id	$\bar{m_k}$	**Top-5 products with sorted highest weights (B_{kd}) associated**	**Repr. countries ($\bar{J}_n$)**
F1	18.27	Misc. Animal Oils (0.78), Bovine and Equine Entrails (0.72), Bovine meat (0.68), Preserved Milk (0.63), Equine (0.62)	Paraguay (2.00)
F2	21.39	Synthetic Woven Fabrics (0.74), Non-retail Synthetic Yarn (0.60), Woven Fabric of less than 85% Discontinuous Synthetic Fibres (0.60), Woven Fabrics of More Than 85% Discontinuous Synthetic Fiber (0.58), Yarn of Less Than 85% Synthetic Fibers (0.53)	United Arab Emirates (2.82)
F3	14.87	Parts of Metalworking Machine Tools (0.74), Interchangeable Tool Parts (0.72), Polishing Stones (0.69), Tool Holders (0.66), Misc. Metalworking Machine-Tools (0.54)	Israel (5.97)
F4	18.67	Aldehyde, Ketone and Quinone-Function Compounds (0.68), Glycosides and Vaccines (0.67), Medicaments (0.65), Inorganic Esters (0.64), Cyclic Alcohols (0.62)	Ireland (4.34)
F5	11.04	Synthetic Rubber (0.87), Acrylic Polymers (0.85), Silicones (0.76), Misc. Polymerization Products (0.71), Tinned Sheets (0.65)	North Korea (3.99)
F6	21.95	Measuring Controlling Instruments (0.61), Mathematical Calculation Instruments (0.59), Misc. Electrical Instruments (0.57), Misc. Heating and Cooling Equipment (0.51), Parts of Office Machines (0.49)	Malaysia (3.00)
F7	31.14	Vehicles Parts and Accessories (0.59), Cars (0.58), Iron Wire (0.53), Trucks and Vans (0.53), Air Pumps and Compressors (0.50)	Belarus (4.20)
F8	33.00	Improved Wood (0.71), Mineral Wool (0.62), Central Heating Equipment (0.62), Aluminium Structures (0.62), Harvesting Machines (0.60)	Belarus (4.20)
F9	16.53	Misc. Electrical Machinery (0.76), Vehicles Stereos (0.72), Misc. Data Processing Equipment (0.64), Video and Sound Recorders (0.57), Calculating Machines (0.55)	Malaysia (3.00)
F10	45.93	Baked Goods (0.67), Metal Containers (0.62), Misc. Edibles (0.59), Misc. Articles of Paper (0.59), Misc. Organic Surfactants (0.58)	Costa Rica (2.06)
F11	33.23	Misc. Articles of Iron (0.65), Carpentry Wood (0.61), Misc. Manufactured Wood Articles (0.60), Sawn Wood Less Than 5mm Thick (0.56), Electric Current (0.51)	Russia (2.93)
F12	38.67	Vegetables (0.60), Fruit or Vegetable Juices (0.54), Misc. Fruit (0.50), Frozen Vegetables (0.48), Apples (0.47)	Peru (2.00)
F13	23.29	Misc. Pumps (0.51), Ash and Residues (0.45), Chemical Wood Pulp of sulphite (0.44), Rolls of Paper (0.43), Worked Nickel (0.43)	Russia (2.93)
F14	46.11	Synthetic Knitted Undergarments (0.76), Misc. Feminine Outerwear (0.74), Misc. Knitted Outerwear (0.73), Men's Shirts (0.70), Blouses (0.67)	Sri Lanka (2.00)
F15	32.12	Misc. Rotating Electric Plant Parts (0.66), Control Instruments of Gas or Liquid (0.58), Valves (0.57), Misc. Rubber (0.56), Misc. Articles of Plastic (0.55)	Philippines (4.01)

Table 6.4: **Complete list of latent features found by the 3RBeP-PFA model.** From left to right, $\bar{m_k}$ is the averaged number of countries having latent feature k active, we list the top-5 products with highest weights B_{kd}; a *representative country* is the country that has the least number of active features among those possessing feature k. $\bar{J}_n$ is the averaged number of active features for each representative country n.

bilities $\bar{J}_n$, are also present in less countries, i.e. $\bar{m_k}$ is low. This observation is in line with methods proposed in [87, 196], as it suggests that capabilities have different degrees of complexity, which plays an important role in the production of goods and services.

Table 6.6 provides a qualitative comparison of our 3RBeP-PFA model against the BeP-PFA and sBeP-PFA models in terms of interpretability strength. For each method to be compared against, we chose the most similar feature across the 10-folds in terms of the Jaccard index. We report the top-25 products and their corresponding factor weights. Compared to all other methods, our model is able to give much shorter and concise descriptions, as weights decrease at a faster pace and the largest weights are significantly larger. Moreover, products in the BeP-PFA list are very heterogeneous. The list for the sBeP-PFA model includes items from a mixture of farming and technological elements, whereas the 3RBeP-PFA list is more homogeneous. All these remarks apply for all the other latent features that are listed in Table 6.4, other comparative tables can be found in the appendix of this chapter. We conclude that our approach enhances interpretability of the latent factors in terms of both conciseness and precision.

Id	**Top-5 products with sorted highest weights (B_{kd}) associated**
F1	Misc. Non-Ferrous Ores (0.40), Petroleum Gases (0.40), Misc. Textile Articles (0.37), Zinc Ore (0.32), Misc. Bituminous Mixtures (0.31)
F2	Sound Recording Media (0.38), Asbestos Products (0.38), Potatoes (0.37), Silver (0.35), Pig Meat (0.32)
F3	Thin Iron Sheets (0.42), Misc. Food-Processing Machinery (0.41), Baked Goods (0.41), Misc. Animal Entrails (0.34), Basketwork (0.34)
F4	Perfumery and Cosmetics (0.45), Misc. Gas Turbines (0.38), Cut Paper (0.35), Misc. Cereal Grains (0.33), Herbicides (0.32)
F5	Bovine (0.49), Misc. Refrigeration Equipment (0.43), Radioactive Chemicals (0.41), Blocks of Iron and Steel (0.41), Rape Seeds (0.40)
F6	Wheat Flour (0.34), Iron and Steel Forging (0.29), Printing Ink (0.29), Waste Paper (0.28), Aluminum (0.26)
F7	Misc. Oil Seeds and Fruits (0.47), Bones, Ivory and Horns (0.44), Temporarily Preserved Fruit (0.43), Cotton Seed Oil (0.42), Inorganic Bases (0.39)
F8	Prepared Explosives (0.48), Confectionary Sugar (0.39), Cigarretes (0.38), Coke (0.37), Misc. Hides and Skins (0.34)
F9	Fish, preserved (0.44), Fresh Fish (0.43), Misc. Animal Origin Materials (0.40), Oranges (0.37), Sheep and Goat Meat (0.37)
F10	Wood and Animal Hair Waste (0.46), Misc. Carpets (0.42), Wool Carpets (0.41), Wool Yarn (0.40), Degreased Sheep Wool (0.38)
F11	Tin (0.41), Vehicles Stereos (0.40), Copper (0.36), Misc. Articles of Paper (0.36), Petroleum Gases (0.36)
F12	Gypsum and Other Calcareous Stone (0.42), Sausage (0.34), Special Products of Textile (0.32), Movie Cameras and Equipment (0.30), Iron Shapes (0.29)
F13	Cigarretes (0.50), Worked Tin and Alloys (0.43), Aluminum (0.38), Bicycles (0.38), Raw Sheep Skin without Wool (0.38)
F14	Precious Metal Ores (0.50), Gold (0.48), Diamonds (0.47), Unmounted Precious Stones (0.43), Electrical Transformers (0.38)
F15	Sulphur (0.40), Fuel Wood and Charcoal (0.34), Misc. Unmilled Cereals (0.33), Household Refrigeration (0.33), Decorative Wood (0.33)

Table 6.5: **Top-15 latent features inferred using the SVD.**

BeP-PFA	SBeP-PFA	3RBeP-PFA
Confectionary Sugar (0.45)	Bovine (0.53)	Miscellaneous Animal Oils (0.78)
Plastic Storage Containers (0.43)	Improved Wood (0.51)	Bovine and Equine Entrails (0.72)
Baked Goods (0.41)	Miscellaneous Vegetable Oils (0.50)	Bovine meat (0.68)
Tissue Paper (0.40)	Butter (0.50)	Preserved Milk (0.63)
Metal Containers (0.39)	Rape Seeds (0.47)	Equine (0.62)
Soaps (0.39)	Miscellaneous Wheat (0.45)	Butter (0.58)
Waste of Man-Made Fibres (0.38)	Pulpwood (0.45)	Misc. Animal Origin Mate. (0.57)
Misc. Organic Surfactants (0.35)	Harvesting Machines (0.45)	Glues (0.56)
Misc. Non-Iron Waste (0.35)	Soil Preparation Machinery (0.44)	Pig Meat (0.53)
Notebooks (0.34)	Misc. Prepared Meats (0.43)	Horse Meat (0.52)
Hydrogenated Oils (0.34)	Bovine meat (0.43)	Malt Extract (0.44)
Iron Structures (0.34)	Coniferous Wood (0.42)	Hay (0.43)
Household Refrigeration (0.34)	Preserved Milk (0.42)	Meat and fish extract (0.35)
Synthetic Knitted Undergarments (0.34)	Misc. Animal Oils (0.41)	Misc. Wheat (0.34)
Chocolate (0.33)	Malt (0.41)	Tractor Units (0.31)

Table 6.6: **Qualitative Evaluation of Topic Interpretability.** We compare the BeP-PFA, sBeP-PFA, and 3RBeP-PFA model. Comparison for any other feature can be found in the Supplementary.

Features correlation. In order to analyze the existing correlations between the inferred factors, we apply our 3RBeP-PFA on the inferred matrix $\mathbf{Z}$ as input data. Such a deep structure, i.e. using two-layer IBP, has already been explored in [40]. As before, we use a bias term, denoted by M-F0, which is active for all countries. In addition to M-F0, our approach extracts another meta-feature, M-F1, which is active for 46.19 countries on average (the list of countries can be found in Table 6.9). The countries with an active M-F1 are those that have more active features in the original model and a larger gross domestic product (GDP). Table 6.8 shows both meta-features, M-F0 and M-F1, which assign different weights to each latent feature from the first layer. For the reader's convenience, we reproduce a compressed version of Table 6.4 in Table 6.7. M-F1 can be

interpreted as the meta-feature that distinguishes between developed countries and developing ones (see Figure 6.3), resulting in a sharp division of the world in terms of capabilities. More importantly these two meta-features divide the features from the original model into three disjoint sets.

The first set contains the latent features whose weight is either zero or insignificant in M-F1: F0, F1, F2, F9, F12 and F14 (highlighted in red). These are the features that define countries with least capabilities, dealing with less complex products like farming or textile (see Table 6.4). It makes sense that more developed countries do not have stronger weights than developing ones for these features. Developed countries might have a non-zero RCA, but they are not exploiting such capabilities better than developing ones.

The second set is composed by F10 (in green), which has a high value in both meta-features. This feature is traded by both developing and developed countries. But the developed countries do trade them more efficiently than developing ones. We can understand these products as those in the capability frontier. We should expect that, as developing countries improve in the future, the weight for F10 would drop towards zero in M-F1.

The last set includes the remaining features (in black). In this case, their weights in M-F0 are negligible compared to their weights in M-F1. These features contain products like chemicals and complex machinery, which are only traded by developed countries. Developing countries have not acquired such capabilities to trade them yet, or they are in the process to do so but are still far behind. Such sharp division among features suggests the existence of a "poverty" or "quiescence trap" in the spirit of [82], a trap of development stasis in which some countries get stuck due to the inability to "acquire" capabilities associated with the production of more complex products.

6.5.2 Time-varying Scenario

To give a further intuition of the interpretability of our time-varying approach, here we apply the dBeP-PFA model to the aggregated SITC database between 1964 and 2010. We consider 92 countries, 501 products, and 47 timestamps. To speed up mixing, we initialize the latent features to the values obtained from a preliminary training with the 2010 data, listed in Table 6.4, and learn the feature activation values for all years.

Quantitative Evaluation

We first evaluate the predictive performance and topic interpretability of dBeP-PFA compared to three alternative methods: Poisson Gamma dynamical system (PGDS), thinned Gamma process Poisson factor analysis (tGaP-PFA), and BeP-PFA. The recently introduced PGDS model puts a

Id	Top-3 products with highest weights
F0	non-coniferous wood, cereal residues, non-iron waste
F1	misc. animal oils, bovine Entails, bovine meat
F2	synthetic woven, synth. yarn, woven $<$ 85% synth.
F3	parts metalworking, tool parts, polishing stones
F4	Aldehyde–Ketone, glycosides–vaccines, medicaments
F5	synthetic rubber, acrylic polymers, silicones
F6	measuring instruments, math inst., electrical inst.
F7	vehicles parts, cars, iron wire
F8	improved wood, mineral wool, heating equipment
F9	elect. machinery, vehicles stereos, data processing eq.
F10	baked goods, metal containers, misc. edibles
F11	misc. articles of iron, carpentry wood, wood articles
F12	vegetables, fruit–vegetable juices, misc. fruit
F13	misc. pumps, ash–residues, chemical wood pulp
F14	synth. undergarments, feminine outerwear, men's shirts
F15	misc. rotating, electric plant parts, control inst. of gas

Table 6.7: **Averaged features learned by the 3RBeP-PFA model.** We list the top-3 products with higher weights for each averaged feature.

Id	$\mathbf{B}'_{1k}$	Id	$\mathbf{B}'_{1k}$
F0	1.00	F8	0.69
F14	0.37	F11	0.68
F12	0.32	F15	0.60
F10	0.17	F10	0.59
F2	0.16	F7	0.52
F1	0.14	F6	0.34
F9	0.13	F13	0.32
F13	0.05	F4	0.31
F6	0.04	F3	0.31
F5	0.04	F5	0.14
F4	0.04	F1	0.05
F15	0.04	F9	0.02
F7	0.03	F2	0.01
F8	0.03	F14	0.00
F11	0.02	F0	0.00
F3	0.02	F12	0.00

(a) M-F0 (b) M-F1

Table 6.8: **Averaged meta-features.** These are inferred by applying our S3R-IBP model to $\mathbf{Z}$ as input.

MF-0	MF-1	List of Countries for each activation pattern of the meta-features
1	0	Pakistan, Syria, Chile, Kyrgyzstan, Zimbabwe, Albania, Tanzania, Bahrain, Laos, Botswana, Bolivia, Bangladesh, Kazakhstan, Senegal, Cuba, Zambia, Namibia, Oman, Turkmenistan, Mongolia, Ethiopia, Mozambique, Iran, Ghana, Cote d'Ivoire, Papua New Guinea, Saudi Arabia, Yemen, Sudan, Trinidad and Tobago, Cameroon, Mauritania, Venezuela, Guinea, Azerbaijan, Algeria, Republic of the Congo, Kuwait, Nigeria, Qatar, Gabon, Libya, Iraq, Angola
1	1	Germany, Italy, United States, Japan, France, China, Austria, Czech Republic, Spain, United Kingdom, Belgium, Sweden, Netherlands, Switzerland, Poland, Denmark, Portugal, Hong Kong, India, Slovenia, Finland, Hungary, Thailand, Israel, Turkey, South Korea, Slovakia, Bulgaria, Romania, Croatia, Estonia, Serbia, Canada, Lithuania, Singapore, Mexico, Panama, Ukraine, Latvia, Malaysia, Brazil, Indonesia, Greece, Bosnia and Herzegovina, Tunisia, Lebanon, Ireland, Vietnam, Philippines, Argentina, Belarus, Egypt, South Africa, North Korea, New Zealand, Russia, Uruguay, El Salvador, United Arab Emirates, Norway, Morocco, Sri Lanka, Moldova, Macedonia, Jordan, Colombia, Australia, Kenya, Mauritius, Peru, Guatemala, Uzbekistan, Dominican Republic, Paraguay, Madagascar, Costa Rica, Honduras, Georgia, Ecuador, Nicaragua, Cambodia, Burma

Table 6.9: **Clustering of countries based on our two-layer 3RBeP-PFA model.** We observe a sharp division of the world in two groups, which aligns with the "quiescence trap" hypothesis [82].

Markovian structure over the dictionary instead of the activation of the features [184]. To run PGDS on our data, we flatten our $\mathbf{X} \in N \times D \times T$ matrix $\mathbf{X}$ into a $ND \times T$ matrix. tGaP-PFA is another Bayesian nonparametric model that learns a smooth underlying parametric function for the activation probability of the features: for certain timestamps t, some features are "thinned" (de-activated) from the underlying random measure. This model follows the same spirit as the dependent IBP in [219], where a Gaussian process (GP) is used to tie the activation of each latent feature over time[5]. BeP-PFA refers to the static model described in Section 6.3 using an IBP prior. In this case, we need to

[5]Inference for the dependent IBP is computationally expensive, and has only been proposed for a Gaussian likelihood. The tGaP-PFA works for Poisson likelihood.

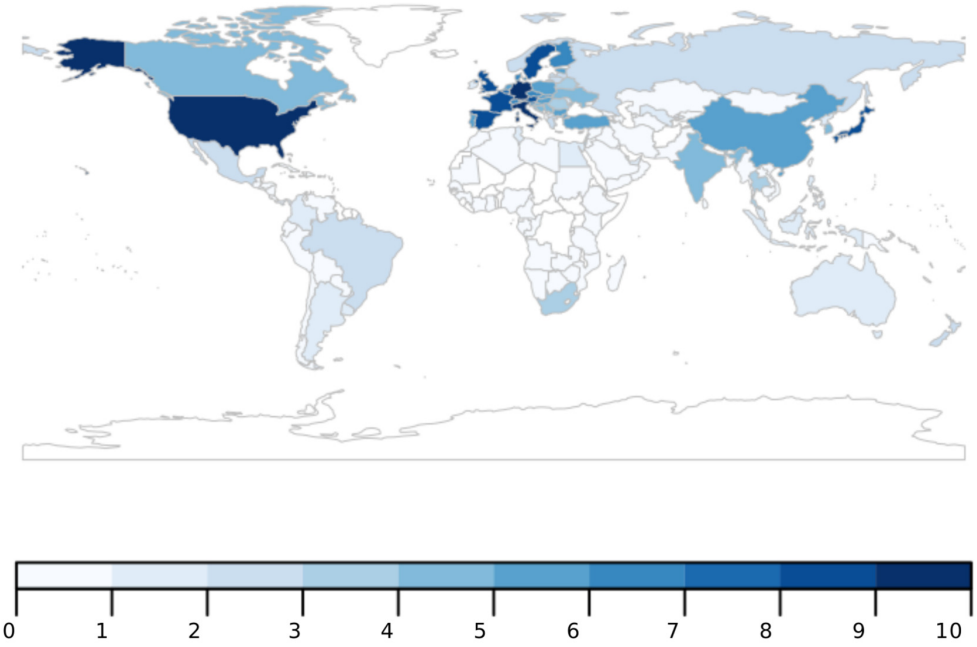


Figure 6.3: **World heat map according to the presence of features associated with meta feature M-F1.** Darker shade indicates presence of more capabilities that are associated with M-F1. Countries with the lightest blue shade only have M-F0 capabilities, whereas there is no data for the countries in white shade.

flatten our input data $\mathbf{X}$ into a $NT \times D$ matrix. The static BeP-PFA model assumes a shared set of latent features over the years, and independent feature activation vectors for each year, e.g., USA in 1965 would be counted as a different country from USA in 1986. Both BeP-PFA and dBeP-PFA assume a bias term (one feature that is active for all countries) to capture global mean effects. For each model, we ran the MCMC inference procedure during 10.000 iterations for ten different splits of the data with 90-10% of the observations in the train and test sets.

Metric	PGDS	tGaP-PFA	BeP-PFA	dBeP-PFA
Log Perplexity	1.912 ± 0.002	1.428 ± 0.001	$\mathbf{1.382 \pm 0.003}$	1.419 ± 0.005
Coherence	-	-469.11 ± 9.562	-506.29 ± 13.470	$\mathbf{-403.70 \pm 31.725}$

Table 6.10: **Quantitative Evaluation of Accuracy and Interpretability for dBeP-PFA.** We compare the following models: PGDS, tGaP-PFA, BeP-PFA and dBeP-PFA in terms of mean and standard error of the test log-perplexities, and topic coherence.

Table 6.10 shows the averaged test log perplexity for each model. Although BeP-PFA is the best model in terms of predictive accuracy, both dBeP-PFA and tGaP-PFA present very close predictive strength. dBeP-PFA is better than BeP-PFA since it imposes more constraints in the solution space

through the Markovian structure. In terms of topic coherence, dBeP-PFA clearly outperforms all the other models, which makes it the best choice to perform a data exploratory analysis of this dataset.

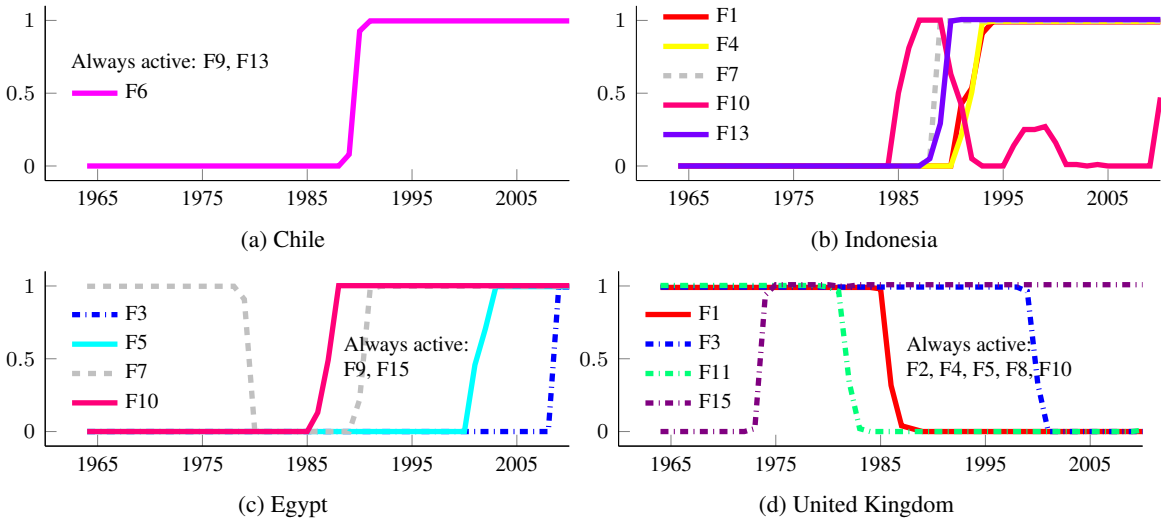
Qualitative Evaluation

Table 6.11 shows the list of inferred latent features using the dBeP-PFA. Compared to the static case, we get less features related to machinery, and more specialized features regarding types of plantations (coffee and sugar are captured by F12 whereas cotton is accounted for by F15). Similar to the static scenario, the induced sparsity in matrix $\mathbf{B}$ makes each latent feature easy to interpret (for instance, F1: objects from a hardware store, F2: chemical products, F3: iron processing, etc...).

Id	$\bar{m}_k$	Top-5 products with highest weights
F0	92	(bias) crude petroleum (0.23), crustaceans (0.18), cereals (0.15), cement (0.14), bones–ivory–horns (0.13)
F1	16.85	light fixtures (0.53), locksmith hardw. (0.46), misc. ceramic ornaments (0.45), bicycles (0.44), misc. manufactured wood articles (0.42)
F2	15.19	inorganic esters (0.54), transmission belts (0.47), chemical products (0.45), nitrogen compounds (0.42), aldehyde compounds (0.42), hormones (0.41)
F3	18.36	iron sheets (0.55), iron wire (0.51), thin iron sheets (0.50), metal cables (0.50), uninsulated steel wire (0.49)
F4	16.89	misc. elect. machinery (0.53), typewriters (0.37), misc. office equipment (0.36), cameras (0.36), calculating machines (0.35)
F5	23.19	soaps (0.53), confectionary sugar (0.44), baked goods (0.42), margarine (0.40), floor polishes (0.35)
F6	18.91	bovine – equine entrails (0.57), bovine meat (0.56), misc. prepared meats (0.51), misc. animal oils (0.48), poultry meat (0.47)
F7	28.89	knit clothing accessories (0.44), linens (0.41), leather accessor (0.40), textile bags (0.39), unbleached cotton woven fabrics (0.36)
F8	15.87	glazes (0.57), textiles fabrics for machinery (0.54), mineral wool (0.53), paper office containers (0.51), misc. mineral materials (0.51)
F9	24.94	misc. vegetables (0.52), grapes and raisins (0.49), misc. fruit (0.48), oranges (0.44), misc. citrus (0.43)
F10	24.77	inorganic bases (0.50), nitrogenous fertilizers (0.44), lubricating petrol. oils (0.40), aluminium (0.34), chemical elements (0.29)
F11	16.09	imitation jewellery (0.53), embroidery (0.46), synth. precious stones (0.44), textile fabrics clothing accessories (0.44), eyewear (0.43)
F12	33.64	coffee (0.71), non-coniferous worked wood (0.42), cane sugar (0.41), cocoa beans (0.37), molasses (0.35)
F13	19.40	copper ores (0.43), chemical wood pulp (0.41), misc. non-ferrous ores (0.40), copper (0.37), zinc ore (0.35)
F14	7.00	pepper (0.69), vegetable plaiting materials (0.68), natural rubber (0.66), unwrought tin and allows (0.58), misc. veg. textile fibres (0.53)
F15	29.94	raw cotton (0.45), cotton linters (0.37), green groundnuts (0.36), misc. animal origin materials (0.34), legumes (0.34)

Table 6.11: **Complete list of latent features found by the dBeP-PFA model.** $\bar{m}_k$ is the averaged number of countries having feature k active across timestamps, we list the top-5 products with highest weights B_{kd}. Orange color corresponds to the latent features represented in Figure 6.5a.

Temporal dynamics. Figure 6.4 shows the feature activation dynamics for four particular examples: Chile, Indonesia, Egypt, and United Kingdom. The three first countries were all able to increase their number of active features over the years, although their dynamics can be attributed to different economic factors. The latent features uncover the internal situation of the countries' economic system at different timestamps. In particular, Chile's main strengths rely on natural resources (F9:



(a) Chile (b) Indonesia

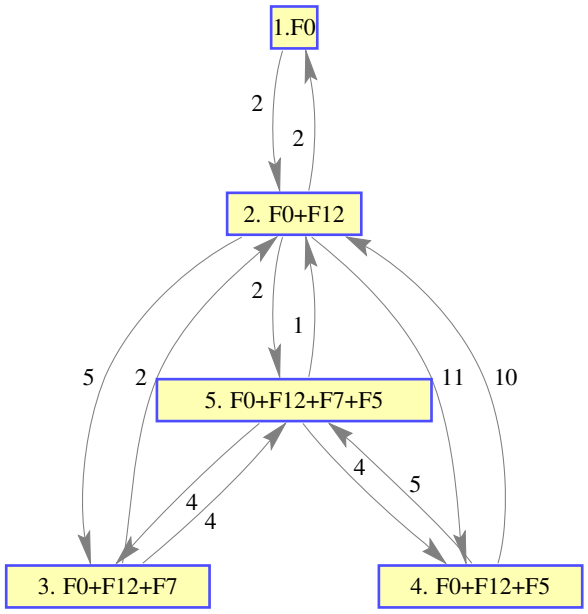
(c) Egypt (d) United Kingdom

Figure 6.4: **Probability of features activation over time using the dynamic dBeP-PFA model.** We have selected four random examples among all countries. We only depict features that are active at some point in time.

vegetables and fruits, F13: ores, ...) and more recently farming industry (acquisition of F6 in the 90s), whereas the growth of Indonesia is mainly due to acquisition of features related to clothing and electronic products. Interestingly, Chile is a well known natural resource producer [83], while the start of Indonesia's growth coincides with the period when the country opened its economy and got an influx of foreign direct investments[133].

Egypt is a country that has very unique dynamics, as shown in Figure 6.4c. There is a sudden fall in the activity of feature F7 (associated to clothing) at the end of the 1970s, after which the number of active features remained steady at its minimum, until 1985 and the beginning of the 1990s. The 80s correspond to years of political and economic unstability [24]. In that period, the country lost most of its export RCA. At the beginning of the second growth, which corresponds to a period of political reforms that made Egypt a more open economy, the country regained the activity in F7, and gradually incorporated F10 (simple chemicals), F5 (basic manufacturing) and F3 (iron processing). The diffusion after the recovery is in accordance with the conclusions from [88] where it was discovered that a country's export basket diffuses over time from only comprising basic products, such as natural resources and clothing products, to also including technology-driven items.

Finally, United Kingdom (UK) is a developed country which has a diversified export portfolio, including chemicals (F2, F10), machinery (F4), and manufacturing (F5, F8) among others. Starting in the 80s, UK loses feature F11 (imitation jewellery, make-up and accessories) and F1 (hardware store items) as a consequence of the de-industrialisation process and structural unemployment during



(a) Subset of the transition model

Transitions	Country	Years
$1 \rightarrow 2$	Congo	1996-1997
$2 \rightarrow 1$	Angola	1976-1977
$2 \rightarrow 3$	Colombia	1977-1978
$2 \rightarrow 4$	Cote D'Ivoire	1998-1999
$2 \rightarrow 4$	Cameroon	1970-1971
$2 \rightarrow 4$	Ecuador	1992-1993
$2 \rightarrow 4$	Honduras	1998-1999
$2 \rightarrow 5$	El Salvador	1985-1986
$3 \rightarrow 2$	Cambodia	1996-1997
$3 \rightarrow 5$	Costa Rica	1998-1999
$4 \rightarrow 2$	Cote D'Ivoire	1983-1984
$4 \rightarrow 2$	Cameroon	1975-1976
$4 \rightarrow 2$	Honduras	1997-1998
$4 \rightarrow 2$	Nicaragua	1982-1983
$4 \rightarrow 5$	Honduras	1999-2000
$5 \rightarrow 3$	Guatemala	1983-1984
$5 \rightarrow 4$	Costa Rica	1999-2000

(b) Transition examples

Figure 6.5: **Transition Model induced by dBeP-PFA.** (a) Subset of the transition model; the label of each node corresponds to the set of active features at this state; edges are weighted by the corresponding number of inferred transition in $\mathbf{Z}$. (b) Transition examples between the nodes depicted in (a).

the Thatcher era [7]. Later in the 2000s, UK also loses F3, corresponding to the delocalization of iron processing industries.

In addition to monitoring the temporal dynamics of each individual country, it is also interesting to study the feature transitions globally, as a simple *transition model*. This is possible given the discrete nature of the feature weights, as features can be either active or inactive. In fact, $\mathbf{Z}_{n\bullet}^{(t)}$ can be interpreted as the *latent state* of country n at timestamp t, which indicates the set of features that are active for country n at time t.

Let $\mathcal{G} = \{\mathcal{M}, \mathcal{E}\}$ be a hidden network, where $\mathcal{M}$ denotes all possible latent states (all possible values for vector $\mathbf{Z}_{n\bullet}^{(t)}$) and $\mathcal{E}$ refers to all directed edges connecting any two elements belonging to $\mathcal{M}$ (full network). For each country n, the sequence $[\mathbf{Z}_{n\bullet}^{(1)}, \mathbf{Z}_{n\bullet}^{(2)}, \ldots, \mathbf{Z}_{n\bullet}^{(T)}]$ corresponds to a certain path within such network. By monitoring which are the most common nodes and most visited edges in the network, we can gain new insights on the developing mechanisms of countries.

Figure 6.5a shows a subset of the transition model induced by the dBeP-PFA model, corresponding to the first steps in countries' development. Node 1 corresponds to the state in which only the bias term is active (no other capabilities), where 16% of the countries are placed each year on average.

From that state, the easiest path is to acquire F12 (coffee, sugar and wood plantations), after which countries might get F7 (basic manufacturing), F5 (clothing) or both. Table 6.5b lists some transition examples corresponding to that particular sub-network. For instance, Honduras is a country with a traditional focus on coffee exports. Based on our model, Honduras manages to diversify its export portfolio by the acquisition of feature F5 and F7 in the 1998-1999 period [25].

6.6 Summary

This chapter presents two extensions of BeP-PFA in order to deal with more flexible structures in the latent space, including structure sparsity and time-dependency. In the static scenario, we have combined the beta-stable process with the R-IBP framework for creating a flexible sparse latent variable prior. We have used this prior together with an additional sparse prior for the weights of the latent features, giving rise to our 3RBeP-PFA model. Our model is a non-negative sparse latent variable model particularly suitable for data exploratory analysis of high-dimensional count data, given its flexibility and interpretability properties. In the time-dependent scenario, we use an mIBP to extend the BeP-PFA. The resulting dBeP-PFA model assumes Markovian structures in the activation of the latent features, which can be interpreted as capabilities required for producing products that a country might acquire or loose over time.

We used these models to explain trade data in a static and time-varying scenarios, e.g., considering the 1964-2010 temporal span. The obtained factors are compact, easy to interpret and show that the number of active capabilities is highly correlated with the diversification and the level of development of an economy. Based on the dynamic approach, we are able to extract a finite state machine of latent patterns that allow us to understand the typical evolution paths of developing countries. The presented approaches are general enough to be directly applicable to any other count-data scenario where data exploration takes precedence over exclusive accuracy, including analysis of gene expression data, topic modeling, or recommendation systems.

7 Conclusions

7.1 Summary

In this thesis, we have constructed tailored Bayesian nonparametric (BNP) models specially suitable for data exploration tasks. We have addressed each new problem from a multidisciplinary point of view, which has been crucial for an appropriate design of the models and their subsequent validation. Globally speaking, the contributions of this thesis can be classified in three key aspects:

1. We have improved *model interpretability via prior and likelihood design*, by including adequate assumptions into the model to better fit reality, using age-gender additive factors in Chapter 3, treatment-specific latent features in Chapter 4, cancer-specific activation weights in Chapter 5, or structured sparsity within the solutions in Chapter 6.

2. We have increased *model flexibility regarding information sharing across observations* by incorporating dependency via Gaussian processes (GPs) in Chapter 3, hierarchical model constructions of Dirichlet processes (DPs), Bernoulli processes (BePs), or Beta processes (BPs) in Chapters 3, 4, and 5, as well as Markovian structures in Chapter 6.

3. We have addressed the issue of *result replicability by combining Bayesian models with statistical methods*, even in small sample size scenarios. We have incorporated significance assessment and multiple hypothesis testing to our approaches, as described in Chapters 4 and 5. The idea of running an unsupervised approach to learn an appropriate latent space for the observations, and then performing statistically significant variable selection for each basis or partition of this latent space is a powerful and novel approach that could bring new insights in a vast amount of applications in medicine, pharmacology, biology or sociology.

More precisely, this doctoral thesis brings actual BNP solutions to non-standard applications in three different scientific fields. Within sport sciences, we have developed the atom-dependent Dirichlet process (ADDP) model, a dependent infinite mixture model for density estimation that was applied on marathon running time data, as explained in Chapter 3. By combining a DP with multiple GPs, we were able to study the impact of age, gender and environment on runner performance [159]. A further analysis based on the hierarchical Dirichlet process (HDP) revealed the most common running patterns along time for the New York City marathon. A direct application of this work results in a fair grading system that is able to compare all runners regardless of their gender or age; such system could be adopted right away in any athletic competition to standardize entry requirements or to grant appropriate rewards among athletes. We also highlighted the existing link of the proposed model with the literature of infinite mixture of experts (IMoE), and thus, we reformulated it as a flexible approach to non-linear regression in general settings with heteroscedastic noise, non-Gaussian likelihoods, or multi-modal distributions.

In the context of cancer research, we have developed BNP approaches with shared and specific components for biomarker discovery and clinico-genetic association in Chapters 4 and 5 respectively. These works have brought novel insights on cancer, and shed light on the mechanisms of action of Codrituzumab, a promising immunotherapy treatment for patients with hepatocellular carcinoma (HCC). An attractive property of the proposed models refers to their capacity to deal with patient heterogeneity, which makes it possible to uncover novel information that previous studies were not able to deliver. The case-control Indian buffet process (C-IBP) or hierarchical Poisson factor analysis (H-PFA) models can be directly applied to any other clinical trial data or genetic association study, respectively. Also, in order to facilitate data exploratory analyses of such kind, we have released open source software, namely the general latent feature model (GLFM) software package that is able to deal with different types of data jointly, together with noise and missing observations. The code is user-friendly, fast, and well documented, such that it can be used straightforwardly by researchers

from other disciplines.

Finally, in the context of international trade, we have proposed two novel doubly-sparse Poisson factor analysis (PFA) models specially suitable for high-dimensional count data. We have investigated the empirical gain of incorporating enhanced prior distributions over the latent observations. In the static scenario, by incorporating the stable-BP within the restricted Indian buffet process (R-IBP), we have allowed for a broader spectrum of sparse solutions that reflect different types of reality. In the dynamic scenario, we have introduced Markovian dynamics in the activation of the latent features based on multiple Markov Indian buffet processes (mIBPs). The obtained time-varying features can be directly interpreted as capabilities necessary for the production of certain products, which countries might acquire or lose over time. The inclusion of mIBP priors results in much smoother and robust transitions, opening the door to a better understanding of the available development paths of countries' economies over time. The proposed models have allowed us to unfold the productive structure of countries via the analysis of their export portfolios, and can be directly used for policy recommendation or future predictions.

Along this thesis, we have developed Markov chain Monte Carlo (MCMC) algorithms based on Gibbs sampling, Metropolis-Hasting steps, slice sampling, forward-filtering backward-sampling (FFBS) procedures, and the help of auxiliary variables to handle non-conjugacies arising into the models. We strongly believe in reproducible research, and thus, code related to most projects has been made publicly available online,[1] including the GLFM toolbox which can be downloaded from Github.[2]

7.2 Future Work

Despite their appealing properties, the impact of BNP models has remained limited so far mostly due to inadequate model assumptions that do not fit reality, or due to unfeasible inference scenarios (e.g., non-conjugate models, slow mixing, multi-modal settings with prohibitive computational cost). Replicability guarantees for Bayesian models are lacking to some extent, particularly in more challenging situations such as very high-dimensional settings and/or small sample size scenarios. Also, understanding the requirements of a particular application and being able to translate them into appropriate model assumptions remains more of an art than a science, but it is nonetheless unavoidable in order to get the utmost out of data. If we manage to alleviate such limitations, there is an incredible

[1] See `www.melaniefpradier.work`

[2] `https://github.com/ivaleraM/GLFM`

opportunity to boost mankind's knowledge in most scientific disciplines, by squeezing huge amounts of available data into more substantial information.

This thesis begets more questions than answers. Our work suggests several paths for further research, both in the technical and application sides. The future lines of research considered herein can be classified in the following categories:

- increasing the flexibility of BNP models to handle more general assumptions or types of data (time-varying streams of data, non-exchangeable observations, complex structures like trees or graphs, etc). In this regard, deep models are promising to address limited likelihood expressivity [170].
- improving inference methods to either better explore the posterior distribution, or scale inference in terms of memory and speed, to allow learning in very high-dimensional or large sample size datasets [11].
- investigating model validation strategies for data exploration tasks and defining appropriate proxies or metrics to better assess model utility in addition to predictive accuracy [41]. On top of empirical evidence, further theoretical analyses are imperatively required in the future.

We provide below a concrete list of ideas for future lines of research.

Generalized ADDP model. The ADDP model of Chapter 3 could be generalized in several ways:

- *multi-dimensional covariates*: we could incorporate an arbitrary number of covariates (nationality, weight, occupation, city, etc), together with an automatic way to select the most promising characteristics. This could be done by using an automatic relevant discrimination (ARD) kernel in the GP.

- *multi-output scenarios*: More output variables could be considered (e.g., intermediate running times, triathlon metrics, etc.). This extension is challenging due to the curse of dimensionality. Some starting points would be the linear model of coregionalization (LMC) or intrinsic coregionalization model [23, 8].

- *other density estimation applications requiring fairness constraints*: the ADDP model could be directly applied to many other comparative studies, such as salary distribution in male and female populations across working sectors, or evolution of height and weight in infants to derive well-calibrated curves for pediatrics. Other applications arise in epidemiological studies, such as monitoring the distribution of scientific output (e.g., number of citations) with respect to age, gender, field of research, etc.

Atom-dependent latent feature model. A promising line of research would be to extend the ADDP model of Chapter 3 to a latent factor scenario, leading to an atom-dependent Indian buffet process model. Dependency could be introduced by assuming a GP prior over each time-varying latent feature. This would allow us to capture elements in a video record that move along time, in the same spirit as in [189]. Inference might suffer from slow mixing, so we might want to consider split and merge moves or gradient-based strategies, such as Hamiltonian Monte Carlo [45, 217].

Another motivating example arises from the clinical trial for Codrituzumab described in Chapter 4: in that study, patients reached different degrees of drug exposure w_n (i.e., this can be interpreted as how much the drug was "absorbed" by each patient n). It would thus be interesting to incorporate dependencies in the features based on the reached level of drug exposure, as follows:

$$x_{nd} \sim \mathcal{N}(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}(w_n), \sigma_x^2), \qquad \text{where } B_{kd}(w_n) = a_{kd}w_n + c_{kd}, \tag{7.1}$$

where the coefficient a_{kd} would encode the direction of change for biomarker d and feature k depending on the drug exposure w_n. Such model would potentially output stronger statistical findings, and could be understood as "correcting" for drug exposure as a confounder effect.

Atom-dependency could also be combined with Markovian structures. For instance, the dynamic Bernoulli process Poisson factor analysis (dBeP-PFA) from Chapter 6 could be improved by additionally allowing a small variation of the latent features over time. Such time-varying drift of the capabilities could be interpreted as historical evolution of export strategies or technologies. For instance, a high-tech capability would present increasing weights for laptops and cellphones in the 90s. In this particular case, we could use the recently introduced Gamma-Poisson auto-regressive chains, which allow for tractable inference in Gamma-distributed Markov chains [3].

Supervised latent feature model. Concerning the C-IBP or the GLFM described in Chapter 4, it would be interesting to include some supervised component to guide the data exploration task in the same spirit as in [163]. Specially in high-dimensional scenarios, there might be multiple local minima with similar predictive performance, but not all the solutions might be as informative with respect to some patient metric. As an example, the latent features could be forced to both explain the observed attributes of patients while being discriminant for progression free survival (PFS).

Efficient approximate inference methods for scalable BNP models. Current BNP models are limited to a moderate number of observations. Recent advances in both MCMC and variational inference methods work with subsets of data in each iteration, which allows scaling up the algorithms

at the expense of a certain accuracy loss in the posterior approximation [11]. A clear promising research path concerns the implementation of efficient Bayesian nonparametric models for big data scenarios. Smart data-driven proposals and efficient moves to speed-up mixing will certainly be needed. BNP models will be much more useful in the future if we manage to develop new inference methods to explore the posterior distribution more efficiently. This research would benefit healthcare applications enormously, since it would render the exploration of vast amounts of electronic health records (EHRs) possible. For example, the considered models in Chapter 5 for genetic association studies would be most useful if these could scale to the hundreds of thousands of dimensions. We could envisage a hybrid variational-MCMC inference scheme, or formulate a black-box variational inference algorithm where non-conjugacy is not problematic anymore [169]. The discrete nature of all the latent variables renders such task particularly challenging in that application, prone to be stuck in local minima. Also, in Chapter 6, the proposed FFBS to infer each Markov chain is relatively slow. Mixing of the MCMC chain could be improved by incorporating a particle Gibbs with ancestor sampling (PGAS) scheme [119]. A starting point could be the work in [212], where the authors propose a dynamic latent feature model using PGAS with Gaussian-distributed latent features. Adapting such model to the Poisson-gamma likelihood-prior case would be useful for data exploratory analyses in count data (because of the non-negativity of the latent features).

Non-linear BNP models for high-dimensional data. Let us imagine a scenario in which the number of dimensions D is orders of magnitude bigger than the number of observations N. To analyze such data, it is a common practice to either select a subset of dimensions, or eventually apply a simple dimensionality reduction scheme as a pre-processing step. A motivating application lies in the field of population genetics where we seek to identify main populations of individuals (centro-european, african, caucasian, etc.) based on vast amounts of genetic variants. A typical scenario would involve thousand of individuals, each of them defined by half a million of genetic variants. Population structure is a serious problem in genome-wide association study (GWAS) that can lead to high false discovery rate (FDR) if not modeled properly. In the future, we plan to extend BNP models to analyze data in scenarios where $D \gg N$. This could be achieved by combining a standard Chinese restaurant process (CRP) or Indian buffet process (IBP) with a non-linear Bayesian dimensionality reduction method such as Gaussian process latent variable model [34]. Such a model would jointly learn the best dimensionality for the reduced space together with the potentially unbounded number of clusters or features in such transformed space.

Nonparametric processes via flexible random measures. Another promising area concerns non-exchangeable priors for dependent Bayesian nonparametric models [54]. So far in this thesis, dependency has been included in the models via smooth variation of the atom locations in the stick-breaking representation in Chapter 3, convex combinations of multiple random measures 4, or hierarchical structures (e.g., Chapter 6), but incorporating dependency at the random measure level is a promising recent area of interest. In fact, novel non-exchangeable BNP models can be obtained by understanding the fundamental properties of the underlying random measures [106, 105]. New completely random measure representations and connections might bring light into improved inference algorithms [146, 118]. Models such as correlated random measures [168] or dependent beta processes [189] illustrate the potential of this line of research. In particular, it would be interesting to introduce dependency over the atom weights in the stick breaking representation, leading to a powerful infinite mixture of GP experts whose weights are input-dependent.

7.3 Discussion

In the last decade, machine learning (ML) systems have become more ubiquitous in complex applications that directly impacts our everyday life. So far, most of these systems have been evaluated in terms of accuracy (e.g., maximization of the log likelihood, minimization of a risk or loss function, etc), but depending on the application at hand, accuracy exclusively might not reflect the actual needs or desired performance of the system. Given the current predominance of accuracy metrics for evaluation in the ML community, the risk exists that we might be "*optimizing too much for the wrong things*[3]".

Nowadays, ML systems are expected not only to have high predictive accuracy, but also to match important criteria such as safety, fairness, or scientific understanding. However, in contrast to measures of performance accuracy, these criteria often cannot be completely quantified, and thus rely on the so-called criterion of *interpretability*: if the system can explain its reasoning, we then can verify whether that reasoning is sound with respect to these auxiliary criteria [41]. "Interpret" means to explain or to present in understandable terms. In the context of ML systems, interpretability refers to the ability to explain or to present in understandable terms to a human (this can be a patient or an expert in the field). Actually, the need for interpretable models is urgent: recent European Union regulation will require algorithms that make decisions based on user-level predictors which "significantly affect" users to provide explanations by 2018, which is commonly refer to as the "right to

[3]F. Doshi-Velez in Talks at Google (April 25th, 2017): "Roadmap for the Rigorous Science of Interpretability".

explanation" [66].

The following question arises: can interpretability be somehow automatized in ML systems? Despite the huge need for interpretable models, data exploration remains quicksand land for the ML community due to a lack of rigorous definitions, evaluation metrics and benchmarks for comparisons. An effort to formalize model assumptions behind each study is needed if we want to speed up data exploratory analyses. Also, translating the requirements from the domain knowledge experts into actual model elements needs special care. Multidisciplinary research efforts will certainly remain crucial, but if we are able to build an *encyclopedia of interpretability requirements and model assumptions*, we might be able to identify similar aspects across applications that might help accelerate and automatize certain types of data exploration.

Current evaluation of interpretability falls in two categories, either application-driven or via a quantifiable-proxy such as sparsity. In either way, we need to put more attention on which assumptions are the most sensitive for each problem at hand, since applications might have different interpretability needs. As stated in [41], the "*need for interpretability stems from an incompleteness in the problem formalization, creating a fundamental barrier to optimization and evaluation*." For example, there might be different kinds of sparsity, each of them suitable for a specific kind of application. Thus, many open questions remain:

- Should a model be sparse in feature weights or sparse in dictionary elements? Should we have sparsity in input or output dimensions?
- Do all applications have the same interpretability needs?
- Where is the fundamental incompleteness in each particular application?
- How can we incorporate human domain-expert knowledge into the model?
- Is there a trade-off between interpretability and accuracy?

In order to enhance model interpretability, we should encode appropriately our assumptions concerning three aspects: the prior, the generative process (likelihood), and the desired output or systems' utility, i.e., in which sense the obtained information will be useful. The assumptions should be as simple as possible (following Occam's Razor principle), and be aligned with reality, e.g., predictions should be consistent with data. Given two models A and B with similar predictive accuracy, we should pick up that one with lower entropy (or minimum description length [73]), or that one which better aligns with expert knowledge. In general, a model will be more interpretable if it contains simple functional components (for instance, additive effects), a clear dependency graph of the involved latent variables which might be identified with specific elements from reality, and a meaningful output design, i.e., clear encoding of the type of information that we seek.

	Marathon (Chapter 3)	Clinical Trial (Chapter 4)	Genetic Associations (Chapter 5)	International Trade (Chapter 6)
How to set up the input (priors)	expert knowledge (humans cannot run a marathon in less than 2h)	normalization + standardized priors	sample hyperparameters + appropriate support (Poisson for count data)	
Assumption underlying the generative process (likelihood)	assume shared mixture weights at shifting locations	learn patient subspace and additional latent dimensions linked to treatment	shared and specific associations across cancer types (hierarchical or partitioned basis)	flexible sparsity assumptions on the latent features (power-law, restricted, time-varying)
Output format (predictive or marginals)	offsets associated to age, gender; race-dependent multiplicative factor	statistical significance, effect size and direction for each clinico-genetic association or biomarker		sparse latent features, i.e., list of countries capabilities
Usability	knowledge + athletic fair grading system	knowledge useful for oncologists and clinicians to improve diagnosis and personalize treatment		knowledge + policy recommendations, future export predictions

Table 7.1: **Model design for interpretability.** Overview of models' elements within this thesis having an impact on interpretability.

Based on these general elements, Table 7.1 gives a final overview of the different mechanisms that we have used to improve interpretability in each model presented within this thesis. To conclude, this doctoral thesis is an attempt to advance the research of BNP models for data exploratory tasks. In the same way that computers have accelerated scientific progress, it is our belief that models for data exploration which are flexible, interpretable, and robust, might lead to an explosion of life-changing scientific discoveries.

8

Inference Details for Poisson Factor Analysis Models and Extensions

8.1 Poisson Factor Analysis

Poisson factorization models have been successfully applied for recommendation systems [68], topic modeling [67], and analysis of Electronic Health Records among others [84]. Let $\mathbf{X} \in \mathbb{N}^{N \times D}$ be a sparse matrix of count-data observations with N samples and D dimensions. The generative process for each single observation x_{nd} in the standard parametric Poisson factor analysis (PFA) is as follows

$$x_{nd} | \boldsymbol{\theta}_{\boldsymbol{n}\bullet}, \mathbf{B}_{\bullet \boldsymbol{d}} \sim \text{Poisson}(\boldsymbol{\theta}_{\boldsymbol{n}\bullet} \mathbf{B}_{\bullet \boldsymbol{d}}) \tag{8.1}$$

$$\theta_{nk} \sim \text{Gamma}(a, b), \quad B_{kd} \sim \text{Gamma}(c, d), \tag{8.2}$$

where $\boldsymbol{\theta}$ is an $N \times K$ matrix of weights, and $\mathbf{B}$ is a $K \times D$ matrix of hidden factors (sometimes called dictionary). A direct non-parametric extension can be obtained by putting a Gamma process prior over $\boldsymbol{\theta}$ [207].

Inference Algorithm

Direct inference in such models is intractable, but we can easily solve the problem using Markov chain Monte Carlo (MCMC) techniques. For each observation x_{nd}, we introduce the auxiliary variables $x'_{nd,1}, \ldots, x'_{nd,K}$ such that $x_{nd} = \sum_{k=1}^{K} x'_{nd,k}$, and $x'_{nd,k} \sim \text{Poisson}(\theta_{nk} B_{kd})$ for $k = 1, \ldots, K$. Each Poisson count is separated in a sum of Poisson contributions corresponding to each latent factor. Given such auxiliary variables, the model is conditionally conjugate, and a Gibbs sampler can be derived straightforwardly. In particular, we use the following theorem:

Theorem 1 *Let $Y_1, \ldots, Y_n$ be Poisson distributed random variables with rates $\lambda_1, \ldots, \lambda_n$ respectively. Let us define $S = \sum_{n=1}^{N} Y_n$. Then,*

$$\{Y_1, \ldots, Y_n\} | S \sim \text{Multinomial}\left(\left\{ \frac{\lambda_i}{\sum_{n=1}^{N} \lambda_n} \right\}_i, S \right). \tag{8.3}$$

Using Theorem 1, $\boldsymbol{x'_{nd,\bullet}}$ can be sampled from a Multinomial given x_{nd}, $\boldsymbol{\theta_{n\bullet}}$ and $\mathbf{B}_{\bullet d}$. The equations for the complete conditional distributions are as follows:

$$p(\theta_{nk} | \boldsymbol{x'_{n\bullet,k}}, \mathbf{B}_{k\bullet}) \propto \text{Gamma}\left(a + \sum_{d=1}^{D} x'_{nd,k},\ b + \sum_{d=1}^{D} B_{kd} \right) \tag{8.4}$$

$$p(B_{kd} | \boldsymbol{x'_{\bullet d,k}}, \boldsymbol{\theta_{\bullet k}}) \propto \text{Gamma}\left(c + \sum_{n=1}^{N} x'_{nd,k},\ d + \sum_{n=1}^{N} \theta_{nk} \right) \tag{8.5}$$

$$p(\boldsymbol{x'_{nd,\bullet}} | x_{nd}, \boldsymbol{\theta_{n\bullet}}, \mathbf{B}_{\bullet d}) \propto \text{Multinomial}\left(\left\{ \frac{\theta_{ni} \mathbf{B}_{id}}{\sum_{k=1}^{K} \theta_{nk} B_{kd}} \right\}_i, x_{nd} \right). \tag{8.6}$$

Based on the conditional distributions, a variational inference scheme can also be derived, since exact expectations can be computed in closed-form due to the Poisson-Gamma conjugacy [67].

8.2 Bernoulli Process Poisson Factor Analysis

The generative model for the Bernoulli process Poisson factor analysis (BeP-PFA) is given by:

$$x_{nd} \sim \text{Poisson}(\mathbf{Z}_{n\bullet} \mathbf{B}_{\bullet d}) \tag{8.7}$$

$$\mathbf{Z} \sim \text{IBP}(\alpha) \tag{8.8}$$

$$\mathbf{B}_{kd} \sim \text{Gamma}\Big(\alpha_B, \frac{\mu_B}{\alpha_B}\Big), \tag{8.9}$$

where $\mathbf{Z}$ is a $N \times K$ matrix of binary weights, and $\mathbf{B}$ is a $K \times D$ matrix of non-negative hidden factors. In the following, we propose two MCMC algorithms: a collapsed Gibbs sampler where matrix $\mathbf{B}$ is marginalized out using a Laplace approximation, and an uncollapsed slice sampler version which allows for parallel sampling of both the elements in $\mathbf{Z}$ and $\mathbf{B}$ given the auxiliary variables described in Section 8.1.

8.2.1 Collapsed Gibbs Sampler

We first propose a collapsed Gibbs sampler where matrix $\mathbf{B}$ is marginalized out, and we only need to sample the elements of matrix $\mathbf{Z}$. We need to compute its posterior distribution:

$$\begin{aligned} p(z_{nk}|\mathbf{X}, \mathbf{Z}_{\neg nk}) &\propto p(z_{nk}|\mathbf{Z}_{\neg nk}) p(\mathbf{X}|\mathbf{Z}) \\ &\propto p(z_{nk}|\mathbf{Z}_{\neg nk}) \int p(\mathbf{X}|\mathbf{Z}, \mathbf{B}) p(\mathbf{B}) d\mathbf{B} \qquad (8.10) \\ &\propto p(z_{nk}|\mathbf{Z}_{\neg nk}) \prod_{d=1}^{D} \int \left(\prod_{n=1}^{N} p(x_{nd}|\mathbf{Z}_{n\bullet}, \mathbf{B}_{\bullet d}) \right) p(\mathbf{B}_{\bullet d}) d\mathbf{B}_{\bullet d} \qquad (8.11) \end{aligned}$$

In order to approximate the integral in (8.11), we resort to a Laplace approximation, which assumes that:

$$\int e^{\psi(\mathbf{B}_{\bullet d})} d\mathbf{B}_{\bullet d} \qquad (8.12)$$

has a peak at a certain value of $\mathbf{B}_{\bullet d}^{\text{MAP}}$. The idea is to Taylor-expand the un-normalized log-posterior of $\mathbf{B}_{\bullet d}$ and approximate $e^{\psi(\mathbf{B}_{\bullet d})}$ by an unnormalized Gaussian. The integral thus corresponds to the normalizing constant of this Gaussian, in our case:

$$\int \left(\prod_{n=1}^{N} p(x_{nd}|\mathbf{Z}_{n\bullet}, \mathbf{B}_{\bullet d}) \right) p(\mathbf{B}_{\bullet d}) d\mathbf{B}_{\bullet d} = e^{\psi(\mathbf{B}_{\bullet d}^{\text{MAP}})} \sqrt{\frac{(2\pi)^K}{|-\nabla\nabla\psi(\mathbf{B}_{\bullet d}^{\text{MAP}})|}} \qquad (8.13)$$

Equations to find maximum a posteriori $\mathbf{B}_{\bullet d}^{\text{MAP}}$. Let us define $\psi(\mathbf{B}_{\bullet d})$ as the un-normalized log-posterior of $\mathbf{B}_{\bullet d}$, i.e,

$$\psi(\mathbf{B}_{\bullet d}) = \sum_{n=1}^{N} \log p(x_{nd}|\mathbf{Z}_{n\bullet}, \mathbf{B}_{\bullet d}) + \log p(\mathbf{B}_{\bullet d}) \qquad (8.14)$$

$$\psi(\mathbf{B}_{\bullet d}) = \sum_{n=1}^{N} x_{nd} \log(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}) - \sum_{n=1}^{N} \mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d} + \sum_{k=1}^{K}(\alpha_B - 1)\log \mathbf{B}_{kd} - \frac{\alpha_B}{\mu_B}\sum_{k=1}^{K}\mathbf{B}_{kd} + R \tag{8.15}$$

$$\text{where } R = -\sum_{n=1}^{N} \log x_{nd}! - K\Big(\alpha_B \log \frac{\mu_B}{\alpha_B} + \log \Gamma(\alpha_B)\Big) \tag{8.16}$$

$$\nabla\psi(\mathbf{B}_{\bullet d}) = \sum_{n=1}^{N} x_{nd}\frac{\mathbf{Z}_{n\bullet}}{\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}} - \sum^{N} \mathbf{Z}_{n\bullet} + (\alpha_B - 1)\frac{1}{\mathbf{B}_{\bullet d}} - \frac{\alpha_B}{\mu_B} \tag{8.17}$$

$$\nabla\nabla\psi(\mathbf{B}_{\bullet d}) = -\sum_{n=1}^{N} x_{nd}\frac{\mathbf{Z}_{n\bullet}\mathbf{Z}_{n\bullet}^T}{(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d})^2} - (\frac{\alpha_B - 1}{\mathbf{B}_{\bullet d}^2})^T\mathbf{I} \tag{8.18}$$

In order to find the maximum value $\mathbf{B}_{\bullet d}^{\text{MAP}}$, we can use either Newton's method or gradient descent. Where applicable, Newton's method might converge faster towards a local maximum or minimum than gradient descent. Newton's method is an iterative method for optimization where each value $\mathbf{B}_{\bullet d}^{(t)}$ at iteration t is computed as:

$$\mathbf{B}_{\bullet d}^{(t)} = \mathbf{B}_{\bullet d}^{(t-1)} + \gamma[\nabla\nabla\psi(\mathbf{B}_{\bullet d}^{(t-1)})]^{-1}\nabla\psi(\mathbf{B}_{\bullet d}^{(t-1)}) \tag{8.19}$$

where $\gamma \in (0, 1]$ is the step-size of the algorithm. Note that for the Laplace approximation to work properly, $-\nabla\nabla\psi(\mathbf{B}_{\bullet d})$ should be a positive semi-definite matrix. This is guaranteed only if $\alpha_B > 1$, so the collapsed Gibbs sampler will only work for shape parameters bigger than one, resulting in non-sparse $\mathbf{B}$ matrices.

8.2.2 Uncollapsed Gibbs Sampler

Inference for the BeP-PFA model can be performed using an uncollapsed Gibbs sampler together with a slice sampler for semi-ordered stick-breaking representation of the Indian buffet process (IBP) [202]. For the sake of completeness, the slice sampling procedure for matrix $\mathbf{Z}$ is described in Algorithm 7. Using the auxiliary random variables described in Section 8.1, the complete conditionals can be easily derived as follows:

$$p(B_{kd}|\mathbf{Z}_{\bullet k}, \boldsymbol{x}'_{\boldsymbol{\bullet d,k}}) \propto \text{Gamma}\left(c + \sum_{n=1}^{N} x'_{nd,k},\ d + \sum_{n=1}^{N} z_{nk}\right) \tag{8.20}$$

$$p(\boldsymbol{x}'_{\boldsymbol{nd,\bullet}}|x_{nd}, \mathbf{B}_{\bullet d}) \propto \text{Multinomial}\left(\left\{\frac{z_{ni}B_{id}}{\sum_{k=1}^{K} z_{nk}B_{kd}}\right\}_{i=1}^{K}, x_{nd}\right) \tag{8.21}$$

Algorithm 7 Slice sampler for the semi-ordered stick-breaking representation of the IBP [202].

1: Sample auxiliary slice variable s for the creation of new sticks, if $s < \mu_{(K^\star)}$, create new sticks using adaptive rejection sampling [65], and sample corresponding feature parameters from prior.

2: Sample $\mathbf{Z}$ matrix. Given the stick weights, each row can be sampled independently and in parallel:

$$p(z_{nk}|rest) \propto \mu_{(k)} \cdot \prod^{D} p(x^{'}_{nd,k}|z_{nk}, \theta_{nk}, B_{kd}). \tag{8.22}$$

3: Remove inactive features.

4: Sample sticks

$$p(\mu_{(k)}|rest) \sim \text{Beta}(m_{\bullet,k}, 1 + N - m_{\bullet,k}), \tag{8.23}$$

where $m_{\bullet,k} = \sum_{i=1}^{N} z_{nk}$.

$$\log p(z_{nk} = 1|\mathbf{Z}_{\neg nk}, \boldsymbol{B}_{\boldsymbol{k}\bullet}, \pi_k) = \frac{1}{1 + \mathrm{e}^{-u_{nk}}}, \tag{8.24}$$

$$\text{where } \; u_{nk} = \sum_{d=1}^{D} x_{nd} \log \left(\frac{\sum_{j \neq k} z_{nj} B_{jd} + B_{kd}}{\sum_{j \neq k} z_{nj} B_{jd}} \right) - \sum_{d=1}^{D} B_{kd} + \log \frac{\pi_k}{1 - \pi_k}. \tag{8.25}$$

8.2.3 Variational Inference

Variational inference is a good option to deal with big and high dimensional scenarios, since it can be scaled up with stochastic extensions, and it has shown good mixing properties in the literature, that allows for a better exploration of the latent space [42]. For the sake of simplicity, we here consider a variational approach that uses a finite Beta-Bernoulli approximation to the IBP.

Let $\mathbf{X}$ be our observations, and $\mathbf{\Phi}$ be the set of hidden variables in the model. The idea behind variational inference is to turn approximate posterior computation into an optimization problem. For that, we choose an easy to handle variational distribution $q(\mathbf{\Phi}|\boldsymbol{\lambda})$ over the set of latent variables $\mathbf{\Phi}$ and optimize its variational free parameters $\boldsymbol{\lambda}$ to approximate the posterior distribution $p(\mathbf{\Phi}|\mathbf{X})$. The objective is to find a variational distribution $q^*(\cdot)$ that minimizes the KL divergence between both distributions, which is equivalent to maximizing the Evidence Lower Bound (ELBO) over the marginal likelihood,

$$q^*(\cdot) = \text{argmax}_{q \in \mathcal{Q}} \mathbb{E}[\log \, p(\mathbf{X}, \mathbf{\Phi}) - \log q(\mathbf{\Phi}|\boldsymbol{\lambda})] \tag{8.26}$$

In our case, we have $\mathbf{\Phi} = \{\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\pi}\}$. Based on the complete conditional distributions derived in Section 8.4.1, exact update equations can be derived for $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, and $\boldsymbol{\beta}$. In particular,

Latent Var	Type	Complete Cond.	Variational Param
π_k	Beta	$g + \sum^N z_{nk}$; $h + N - \sum^N z_{nk}$	$\widetilde{\tau}_{k1}$; $\widetilde{\tau}_{k2}$
z_{nk}	Bernoulli	$1/\left(1 - \exp^{-u_{nk}}\right)$	$\widetilde{\nu}_{nk}$
θ_{nk}	Gamma	$a + z_{nk} \sum^D x^{'}_{nd,k}$; $b + z_{nk} \sum^D \beta_{kd}$	$\widetilde{\theta}^{shp}_{nk}$; $\widetilde{\theta}^{rte}_{nk}$
β_{kd}	Gamma	$c + \sum^N z_{nk} x^{'}_{nd,k}$; $d + \sum^N z_{nk}\theta_{nk}$	$\widetilde{\beta}^{shp}_{kd}$; $\widetilde{\beta}^{rte}_{kd}$
$\boldsymbol{x^{'}_{nd,\cdot}}$	Multinomial	$\log\left(z_{nk}\theta_{nk}\beta_{kd}\right)$, $0 < k < K$	$\widetilde{\boldsymbol{\phi}}_{\boldsymbol{nd},\cdot}$

Table 8.1: Summary and notation for the variational inference procedure in the BeP-PFA.

$$\widetilde{\tau}_{k1} = g + \sum^N \widetilde{\nu}_{nk} \tag{8.27}$$

$$\widetilde{\tau}_{k2} = h + N - \sum^N \widetilde{\nu}_{nk} \tag{8.28}$$

$$\widetilde{\theta}^{shp}_{nk} = a + \widetilde{\nu}_{nk} \sum^D x_{nd} \widetilde{\phi}_{nd,k} \tag{8.29}$$

$$\widetilde{\theta}^{rte}_{nk} = b + \widetilde{\nu}_{nk} \sum^D \frac{\widetilde{\beta}^{shp}_{nd}}{\widetilde{\beta}^{rte}_{nd}} \tag{8.30}$$

$$\widetilde{\beta}^{shp}_{kd} = c + \sum^N \widetilde{\nu}_{nk} x_{nd} \widetilde{\phi}_{nd,k} \tag{8.31}$$

$$\widetilde{\beta}^{rte}_{kd} = d + \sum^N \widetilde{\nu}_{nk} \frac{\widetilde{\theta}^{shp}_{nk}}{\widetilde{\theta}^{rte}_{nk}} \tag{8.32}$$

Update equations for $\mathbf{Z}$ and $\boldsymbol{x^{'}_{nd,\cdot}}$ cannot be derived analytically. Instead, we compute a noisy gradient step based on a black-box procedure [169].

Black-box Variational Inference

Our objective is find the parameters $\boldsymbol{\lambda}$ that maximize the ELBO $\mathcal{L}\left(\boldsymbol{\lambda}\right)$:

$$\mathcal{L}\left(\boldsymbol{\lambda}\right) = \mathbb{E}_q[\log p\left(\mathbf{X}, \boldsymbol{\Phi}\right) - \log q\left(\boldsymbol{\Phi}|\boldsymbol{\lambda}\right)] \tag{8.33}$$

The general form for the ELBO derivative with respect to the free parameters $\boldsymbol{\lambda}$ is given as:

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L} = \mathbb{E}_q[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\Phi}|\boldsymbol{\lambda}) \big(\log p(\mathbf{X}, \boldsymbol{\Phi}) - \log q(\boldsymbol{\Phi}|\boldsymbol{\lambda}) \big)] \tag{8.34}$$

Let $\boldsymbol{\Phi} = \{\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\pi}\}$ be the set of all latent variables in our model, and $\boldsymbol{\lambda}$ all the free parameters to characterize the variational distribution $q\left(\boldsymbol{\Phi}|\boldsymbol{\lambda}\right)$. The joint distribution is given as:

$$\begin{aligned}
\log p\left(\mathbf{X}, \boldsymbol{\Phi}\right) = & \sum^{N}\sum^{D} \log \text{Poisson}\left(x_{nd}; \sum^{K} z_{nk}\theta_{nk}\beta_{kd}\right) \\
& + \sum^{K}\sum^{D} \log \text{Gamma}\left(\beta_{kd}; a_{\beta}, b_{\beta}\right) \\
& + \sum^{N}\sum^{K} \log \text{Bernoulli}\left(z_{nk}; \pi_k\right) \\
& + \sum^{N}\sum^{K} \log \text{Gamma}\left(\theta_{nk}; a_{\theta}, b_{\theta}\right) \\
& + \sum^{K} \log \text{Beta}\left(\pi_k; a_{\pi}, b_{\pi}\right)
\end{aligned} \tag{8.35}$$

Using a mean-field approximation, the whole variational distribution $q(\boldsymbol{\Phi}|\boldsymbol{\lambda})$ factorizes as

$$\begin{aligned}
\log q\left(\boldsymbol{\Phi}|\boldsymbol{\lambda}\right) = & \sum^{N}\sum^{D} \log \text{Multinomial}(\boldsymbol{x'_{nd,\cdot}}|x_{nd}, \widetilde{\phi}_{nd}) \\
& + \sum^{K}\sum^{D} \log \text{Gamma}(\beta_{kd}; \widetilde{\beta}^{shp}_{kd}, \widetilde{\beta}^{rte}_{kd}) \\
& + \sum^{N}\sum^{K} \Big(\log \text{Gamma}(\theta_{nk}; \widetilde{\theta}^{shp}_{nk}, \widetilde{\theta}^{rte}_{nk}) \\
& + \log \text{Bernoulli}\left(z_{nk}; \widetilde{\nu}_{nk}\right) \Big) \\
& + \sum^{K} \log \text{Beta}\left(\pi_k; \widetilde{\tau}_{k1}, \widetilde{\tau}_{k2}\right)
\end{aligned} \tag{8.36}$$

The idea of Black-box Variational Inference is to update each variational free parameter λ_i following a step in the direction of the noisy gradient of the ELBO,

$$\lambda_i^{(t)} = \lambda_i^{(t-1)} + \rho_t \nabla_{\lambda_i} \mathcal{L}(\boldsymbol{\lambda}) \tag{8.37}$$

To reduce the variance of the gradient estimator, we perform rao-blackwellization for each variable, as explained in [169]. Each component of the gradient can then be written as

$$\nabla_{\lambda_i}\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_q\Big[\nabla_{\lambda_i}\log q(\boldsymbol{\Phi}_i|\lambda_i)\big(\log p(\mathbf{X},\boldsymbol{\Phi}_i) - \log q(\boldsymbol{\Phi}_i|\lambda_i)\big)\Big] \tag{8.38}$$

where $\boldsymbol{\Phi}_i$ is the set of latent variables in the Markov blanket of ϕ_i, and ϕ_i is the latent variable whose variational distribution includes λ_i.

Derivation of noisy gradients

- Equations for $\widetilde{\boldsymbol{\phi}}_{\boldsymbol{nd},\cdot}$.

$$\log q(\boldsymbol{x'_{nd,\cdot}}|\widetilde{\phi}_{nd}, x_{nd}) = \sum_{}^{K} x'_{nd,k}\log\widetilde{\phi}_{nd,k} + \log(x_{nd}!) - \sum_{}^{K}\log(x'_{nd,k}!) \tag{8.39}$$

$$\nabla_{\widetilde{\phi}_{nd,k}}\log q(\boldsymbol{x'_{nd,\cdot}}|\widetilde{\phi}_{nd}, x_{nd}) = \frac{x'_{nd,k}}{\widetilde{\phi}_{nd,k}} \tag{8.40}$$

Technical note: Since $\widetilde{\boldsymbol{\phi}}_{\boldsymbol{nd},\cdot}$ are the variational parameters of a multinomial distribution, we need to satisfy $\sum_{k=1}^{K}\widetilde{\phi}_{nd,k} = 1$. For this purpose, we use the softmax function $\mathcal{P}(\boldsymbol{x})_j = \frac{e^{x_j}}{\sum^K e^{x_k}}$ to transform the original unconstrained variational parameters. The equations should be properly modified using the chain rule, as follows:

$$\begin{aligned}\log q(\boldsymbol{x'_{nd,\cdot}}|\widetilde{\phi}_{nd}, x_{nd}) = &\sum_{}^{K} x'_{nd,k}\log\mathcal{P}(\widetilde{\phi}_{nd,k}) \\ &+ \log(x_{nd}!) - \sum_{}^{K}\log(x'_{nd,k}!)\end{aligned} \tag{8.41}$$

$$\begin{aligned}\nabla_{\widetilde{\phi}_{nd,k}}\log q(\boldsymbol{x'_{nd,\cdot}}|\widetilde{\phi}_{nd}, x_{nd}) = \\ &-\sum_{j\neq k}\frac{x'_{nd,j}}{\mathcal{P}(\widetilde{\boldsymbol{\phi}}_{\boldsymbol{nd},\cdot})_j}\frac{e^{\widetilde{\phi}_{nd,j}}e^{\widetilde{\phi}_{nd,k}}}{(\sum_{r=1}^{K}e^{\widetilde{\phi}_{nd,r}})^2} \\ &+\frac{x'_{nd,k}}{\mathcal{P}(\widetilde{\boldsymbol{\phi}}_{\boldsymbol{nd},\cdot})_k}\frac{e^{\widetilde{\phi}_{nd,k}}\sum_{r=1}^{K}e^{\widetilde{\phi}_{nd,r}} - e^{2\widetilde{\phi}_{nd,k}}}{(\sum_{r=1}^{K}e^{\widetilde{\phi}_{nd,r}})^2}\end{aligned} \tag{8.42}$$

Finally, the Markov blanket of each variable $x'_{nd,k}$ is $\mathcal{M}[x'_{nd,k}] = \{x'_{nd,k}, \boldsymbol{x'_{nd,\neg k}}, \boldsymbol{z_{n\cdot}}, \boldsymbol{\theta_{n\cdot}}, \boldsymbol{\beta_{\cdot d}}\}$ and their corresponding probability distribution is given by

$$\begin{aligned}\log p\big(\mathcal{M}[x^{'}_{nd,k}]\big) = \\ \log \text{Mult}\big(\boldsymbol{x^{'}_{nd,\cdot}}|x_{nd}, (\boldsymbol{z_{n\cdot}} \odot \boldsymbol{\theta_{n\cdot}})\boldsymbol{\beta_{\cdot d}}\big) \\ + \sum^{K} \Big(\log \text{Bernoulli}\big(z_{nk}|\pi_k\big) + \log \text{Gamma}\big(\theta_{nk}|\alpha_\theta, \frac{\mu_\theta}{\alpha_\theta}\big) \\ + \log \text{Gamma}\big(\beta_{kd}|\alpha_\beta, \frac{\mu_\beta}{\alpha_\beta}\big)\Big)\end{aligned} \tag{8.43}$$

The gradient coordinates for each free parameter $\widetilde{\phi}_{nd,k}$ is computed by plugging in eq. (8.41), (8.42), and (8.43) into (8.38).

- Equations for $\mathbf{Z}$:

$$\log q(z_{nk}|\widetilde{\nu}_{nk}) = z_{nk} \log \widetilde{\nu}_{nk} + (1 - z_{nk}) \log(1 - \widetilde{\nu}_{nk}) \tag{8.44}$$

$$\nabla_{\widetilde{\nu}_{nk}} \log q(z_{nk}|\widetilde{\nu}_{nk}) = \frac{z_{nk}}{\widetilde{\nu}_{nk}} - \frac{1 - z_{nk}}{1 - \widetilde{\nu}_{nk}} \tag{8.45}$$

Technical note: $\widetilde{\nu}_{nk}$ is the variational parameter of a Bernoulli distribution, so it should belong to the interval $[0, 1]$. We use the sigmoid transformation function $\mathcal{F}(x) = \frac{1}{1+e^{-x}}$. The new equations are the following:

$$\begin{aligned}\log q(z_{nk}|\widetilde{\nu}_{nk}) = z_{nk} \log \mathcal{F}(\widetilde{\nu}_{nk}) \\ + (1 - z_{nk}) \log \big(1 - \mathcal{F}(\widetilde{\nu}_{nk})\big)\end{aligned} \tag{8.46}$$

$$\nabla_{\widetilde{\nu}_{nk}} \log q(z_{nk}|\widetilde{\nu}_{nk}) = \mathcal{F}^{'}(\widetilde{\nu}_{nk}) \left(\frac{z_{nk}}{\mathcal{F}(\widetilde{\nu}_{nk})} - \frac{1 - z_{nk}}{1 - \mathcal{F}(\widetilde{\nu}_{nk})} \right) \tag{8.47}$$

where $\mathcal{F}^{'}(\widetilde{\nu}_{nk}) = \frac{e^{-x}}{(1+e^{-x})^2}$.

The markov blanket for each variable z_{nk} is $\mathcal{M}[z_{nk}] = \{z_{nk}, \pi_k, x^{'}_{nd,k}, \theta_{nk}, \beta_{kd}\}$ and their corresponding probability distribution is given by

$$\begin{aligned}\log p\big(\mathcal{M}[z_{nk}]\big) &= \log \mathrm{Beta}\big(\pi_k|a_\pi, b_\pi\big) + \log \mathrm{Bernoulli}\big(z_{nk}|\pi_k\big) \\ &+ \log \mathrm{Gamma}\big(\theta_{nk}|\alpha_\theta, \frac{\mu_\theta}{\alpha_\theta}\big) + \sum^{D} \log \mathrm{Gamma}\big(\beta_{kd}|\alpha_\beta, \frac{\mu_\beta}{\alpha_\beta}\big) \\ &+ \sum^{D} \log \mathrm{Poisson}\big(x^{'}_{nd,k}|z_{nk}\theta_{nk}\beta_{kd}\big) \end{aligned} \tag{8.48}$$

The gradient estimator is computed by plugging in eqs. (8.46), (8.47), and (8.48) into (8.38).

If we run the algorithm with these equations, all variational parameters for $\mathbf{Z}$ might tend to zero. One explanation for this might be the strong correlation between $\mathbf{Z}$ and $\mathbf{X}^{'}$. To solve it, we use the probability of the Markov blanket $\mathcal{M}[z_{nk}]$ after collapsing the auxiliary variables $\mathbf{X}^{'}$. The new joint probability is

$$\begin{aligned}\log p\big(\mathcal{M}[z_{nk}]\big) &= \log \mathrm{Beta}\big(\pi_k|a_\pi, b_\pi\big) \\ &+ \sum^{K} \Big(\log \mathrm{Bernoulli}\big(z_{nk}|\pi_k\big) + \log \mathrm{Gamma}\big(\theta_{nk}|\alpha_\theta, \frac{\mu_\theta}{\alpha_\theta}\big)\Big) \\ &+ \sum^{K}\sum^{D} \log \mathrm{Gamma}\big(\beta_{kd}|\alpha_\beta, \frac{\mu_\beta}{\alpha_\beta}\big) \\ &+ \sum^{D} \log \mathrm{Poisson}\big(x_{nd}|(\boldsymbol{z_{n\cdot}} \odot \boldsymbol{\theta_{n\cdot}})\boldsymbol{\beta_{\cdot d}}\big) \end{aligned} \tag{8.49}$$

8.3 Spike and Slab Bernoulli Process Poisson Factor Analysis

To get even sparser topics, we may use a spike and slab prior on each element of $\mathbf{B}$ by incorporating a binary matrix $\overline{\boldsymbol{\Theta}}_{\boldsymbol{d}}$ that works as a selection mask on the feature matrix $\mathbf{B}_{\bullet d}$. The generative model in this case is given by:

$$x_{nd} \sim \mathrm{Poisson}\Big(\mathbf{Z}_{n\bullet}\left(\overline{\boldsymbol{\Theta}}_{\boldsymbol{d}} \odot \mathbf{B}_{\bullet d}\right)\Big) \tag{8.50}$$

$$\mathbf{Z} \sim \mathrm{IBP}\left(\alpha\right) \tag{8.51}$$

$$\mathbf{B}_{kd} \sim \mathrm{Gamma}\left(\alpha_B, \frac{\mu_B}{\alpha_B}\right) \tag{8.52}$$

$$\overline{\boldsymbol{\Theta}} \sim \mathrm{IBP}(\alpha_\beta, c_\beta, \sigma_\beta). \tag{8.53}$$

An equivalent formulation can be written as:

$$x_{nd} \sim \text{Poisson}(\mathbf{Z}_{n\bullet}\overline{\mathbf{B}_{\bullet d}}) \tag{8.54}$$

$$\mathbf{Z} \sim \text{IBP}(\alpha) \tag{8.55}$$

$$\overline{\mathbf{B}_{kd}}|\overline{\theta}_{kd} \sim \text{Gamma}\Big(\alpha_B, \frac{\mu_B}{\alpha_B}\Big)^{\overline{\theta}_{kd}} + \delta_0^{(1-\overline{\theta}_{kd})} \tag{8.56}$$

$$\overline{\mathbf{\Theta}} \sim \text{IBP}(\alpha_\beta, c_\beta, \sigma_\beta), \tag{8.57}$$

where $\overline{\mathbf{B}_{kd}} = \mathbf{B}_{kd} \cdot \overline{\theta}_{kd}$, and $\overline{\mathbf{B}_{kd}}$ follows a spike and slab distribution where $\overline{\theta}_{kd}$ selects between the Gamma distribution (slab) or a delta probability mass at zero. A third way to write down the generative model would be as follows:

$$x_{nd} \sim \text{Poisson}(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}) \tag{8.58}$$

$$\mathbf{Z} \sim \text{IBP}(\alpha) \tag{8.59}$$

$$\mathbf{B}_{\bullet d}^T|\overline{\mathbf{\Theta}}_{\boldsymbol{d}} \sim \text{SGP}(\alpha_B, \mu_B, \overline{\mathbf{\Theta}}_{\boldsymbol{d}}) \tag{8.60}$$

$$\overline{\mathbf{\Theta}} \sim \text{IBP}(\alpha_\beta, c_\beta, \sigma_\beta), \tag{8.61}$$

where each column of matrix $\mathbf{B}$ is a draw from a spike and slab Gamma process. Such process is defined as the multiplication of a Gamma process together with a Bernoulli process. We call such process *IBP compound Gamma process*. This approach is very similar in spirit to [13], where they use a tree-parameter Indian buffet process (3P-IBP) over the feature matrix in a Dirichlet-Multinomial topic model. We present here its counterpart with a Poisson-Gamma formulation. Regarding inference, the model is not conjugate anymore because of the spike component in the features, e.g., the binary matrix $\overline{\mathbf{\Theta}}_{\boldsymbol{d}}$. We thus derive a collapsed Gibbs sampler where matrix $\mathbf{B}$ is marginalized out[1].

8.3.1 Collapsed Gibbs Sampler

Since the spike and slab BeP-PFA is not conjugate anymore, we propose a collapsed Gibbs sampler where we integrate matrix $\mathbf{B}$ and only need to sample the two binary matrices $\mathbf{Z}$ and $\overline{\mathbf{\Theta}}$. For that, we will need to compute the conditional probabilities $p(z_{nk}|\mathbf{X}, \mathbf{Z}_{\neg nk}, \overline{\mathbf{\Theta}})$ and $p(\overline{\theta}_{kd}|\mathbf{X}, \overline{\mathbf{\Theta}}_{\neg k,d}, \mathbf{Z})$.

[1] Variational inference in this setting becomes tedious due to the discrete nature of the latent space, resulting in multiple local minima.

Sampling elements of Z given $\overline{\mathbf{\Theta}}$

$$
\begin{aligned}
p(z_{nk}|\mathbf{X}, \mathbf{Z}_{\neg nk}, \overline{\mathbf{\Theta}}) &\propto p(z_{nk}|\mathbf{Z}_{\neg nk}) p(\mathbf{X}|\mathbf{Z}, \overline{\mathbf{\Theta}}) \\
&\propto p(z_{nk}|\mathbf{Z}_{\neg nk}) \int p(\mathbf{X}|\mathbf{Z}, \mathbf{B}, \overline{\mathbf{\Theta}}) p(\mathbf{B}) d\mathbf{B} \qquad (8.62) \\
&\propto p(z_{nk}|\mathbf{Z}_{\neg nk}) \int p(\mathbf{X}|\mathbf{Z}, \overline{\boldsymbol{\beta}}) p(\overline{\boldsymbol{\beta}}) d\overline{\boldsymbol{\beta}} \qquad (8.63) \\
&\propto p(z_{nk}|\mathbf{Z}_{\neg nk}) \prod_{d=1}^{D} \int \left(\prod_{n=1}^{N} p(x_{nd}|\mathbf{Z}_{n\bullet}, \overline{\boldsymbol{\beta}}_{\boldsymbol{d}}) \right) p(\overline{\boldsymbol{\beta}}_{\boldsymbol{d}}) d\overline{\boldsymbol{\beta}}_{\boldsymbol{d}}, \qquad (8.64)
\end{aligned}
$$

where $p(z_{nk} = 1|\mathbf{Z}_{\neg nk}) = \frac{m_k}{N}$, $m_k = |\{n \,|\, z_{nk} = 1\}|$, and the marginal likelihood $p(\mathbf{X}|\mathbf{Z}, \overline{\mathbf{\Theta}})$ can be computed using Laplace approximation, since the function to optimize is always log-concave for $\alpha_B > 1$. Sparsity in the features is obtained through the binary mask $\overline{\mathbf{\Theta}}$.

In order to sample new features, we follow the approach described in [72]. In particular, there are an infinite number of remaining columns which contain all zeros. For any particular feature $k > K$, the probability of activation of any element is zero. Nonetheless, we can sample the number of columns that become nonzero, k_{new}, according to

$$
p(k_{new}) \propto \text{Poisson}\Big(k_{new}; \frac{\alpha}{N}\Big) p(\mathbf{X}|\mathbf{Z}_{new}), \qquad (8.65)
$$

where $\mathbf{Z}_{new}$ is the feature assignment matrix with k_{new} additional columns set to one for observation n, and zero otherwise. We compute these probabilities for $k_{new} = 0, \ldots, K_{max}$ for some K_{max}, normalize and sample from the resulting multinomial.

Sampling elements of $\overline{\mathbf{\Theta}}$ given Z

$$
\begin{aligned}
p(\overline{\theta}_{kd}|\mathbf{X}, \overline{\mathbf{\Theta}}_{\neg k,d}, \mathbf{Z}) &\propto p(\overline{\theta}_{kd}|\overline{\mathbf{\Theta}}_{\neg kd}) p(\mathbf{X}_{\boldsymbol{d}}|\mathbf{Z}, \overline{\mathbf{\Theta}}_{\boldsymbol{d}}) \\
&\propto p(\overline{\theta}_{kd}|\overline{\mathbf{\Theta}}_{\neg k,d}) \int p(\mathbf{X}_{\boldsymbol{d}}|\mathbf{Z}, \overline{\boldsymbol{\beta}}_{\boldsymbol{d}}) p(\overline{\boldsymbol{\beta}}_{\boldsymbol{d}}) d\overline{\boldsymbol{\beta}}_{\boldsymbol{d}} \qquad (8.66) \\
&\propto p(\overline{\theta}_{kd}|\overline{\mathbf{\Theta}}_{\neg k,d}) \int \left(\prod_{n=1}^{N} p(x_{nd}|\mathbf{Z}_{n\bullet}, \overline{\boldsymbol{\beta}}_{\boldsymbol{d}}) \right) p(\overline{\boldsymbol{\beta}}_{\boldsymbol{d}}) d\overline{\boldsymbol{\beta}}_{\boldsymbol{d}}, \qquad (8.67)
\end{aligned}
$$

where $p(\overline{\theta}_{kd} = 1|\overline{\mathbf{\Theta}}_{\neg kd}) = \frac{r_k - \sigma_\beta}{D - 1 + c_\beta}$ and $r_k = |\{d \,|\, \overline{\theta}_{kd} = 1\}|$. Here again, in order to approximate the marginal likelihood $p(\mathbf{X}_{\boldsymbol{d}}|\mathbf{Z}, \overline{\mathbf{\Theta}}_{\boldsymbol{d}})$, we resort to Laplace approximation.

Similar as before, we can sample the number of new features k_{new} that should be activated for observation n as:

$$p(k_{new}) \propto \text{Poisson}(k_{new}; \epsilon) \cdot p(\mathbf{X}|\mathbf{Z}_{new}), \tag{8.68}$$

$$\text{where } \epsilon = \alpha_\beta \frac{\Gamma(1+c_\beta)\Gamma(D-1+c_\beta+\sigma_\beta)}{\Gamma(D+c_\beta)\Gamma(c_\beta+\sigma_\beta)}. \tag{8.69}$$

8.4 Dynamic Bernoulli Process Poisson Factor Analysis

For the sake of completeness, we state here again the equations describing the dynamic Bernoulli process Poisson factor analysis (dBeP-PFA) model:

$$x_{nd}^{(t)} \sim \text{Poisson}\left(\mathbf{Z}_{n\bullet}^{(t)} \mathbf{B}_{\bullet d}\right) \tag{8.70}$$

$$\mathbf{Z}_{n\bullet}^{(\bullet)} \sim \text{mIBP}(\alpha) \tag{8.71}$$

$$\mathbf{B}_{kd} \sim \text{Gamma}\left(\alpha_B, \frac{\mu_B}{\alpha_B}\right), \tag{8.72}$$

where the Markov Indian buffet process (mIBP) prior is equivalent to the following parametric construction as $K \to \infty$:

$$\begin{aligned} a_k &\sim \text{Beta}(\frac{\alpha}{K}, 1), && (8.73)\\ b_k &\sim \text{Beta}(\gamma, \delta), && (8.74)\\ z_{nk}^{(t)}|a_k, b_k &\sim \text{Bernoulli}\left(a_k^{1-z_{nk}^{(t-1)}} b_k^{z_{nk}^{(t-1)}}\right). && (8.75) \end{aligned}$$

The variables α_B, $hypmuB$, α, γ, and δ are the model hyperparameters.

8.4.1 MCMC Inference

The complete conditionals in this case are:

$$p(B_{kd}|\mathbf{Z}_{\bullet \boldsymbol{k}}^{(\bullet)}, \boldsymbol{r}_{\bullet \boldsymbol{d},\boldsymbol{k}}^{(\bullet)}) \propto \text{Gamma}\left(\alpha_B + \sum_{n=1}^{N}\sum_{t=1}^{T} r_{nd,k}^{(t)},\ \frac{\mu_B}{\alpha_B} + \sum_{n=1}^{N}\sum_{t=1}^{T} z_{nk}^{(t)}\right) \tag{8.76}$$

$$p(\boldsymbol{r}_{\boldsymbol{nd},\cdot}^{(\boldsymbol{t})}|x_{nd}^{(t)}, \mathbf{Z}_{n\bullet}^{(t)}, \mathbf{B}_{\bullet d}) \propto \text{Multinomial}\left(\left\{\frac{z_{ni}^{(t)}\mathbf{B}_{id}}{\sum_{k=1}^{K} z_{nk}^{(t)} B_{kd}}\right\}_{i=1,\ldots,K^+}, x_{nd}^{(t)}\right) \tag{8.77}$$

The most challenging part is how to sample $p(\mathbf{Z}|\mathbf{X}, \mathbf{B})$. For each country n and latent feature k, we can write:

$$p(\mathbf{X}_{n\bullet}^{(1:t)}, z_{nk}^{(t)}|-) = p(\mathbf{X}_{n\bullet}^{(t)}|z_{nk}^{(t)}, -)\; p(\mathbf{X}_{n\bullet}^{(1:t-1)}, z_{nk}^{(t)}|-) \tag{8.78}$$

$$= p(\mathbf{X}_{n\bullet}^{(t)}|z_{nk}^{(t)}, -) \sum_{z_{nk}^{(t-1)}} p(\mathbf{X}_{n\bullet}^{(1:t-1)}, z_{nk}^{(t)}, z_{nk}^{(t-1)}|-) \tag{8.79}$$

$$= p(\mathbf{X}_{n\bullet}^{(t)}|z_{nk}^{(t)}, -) \sum_{z_{nk}^{(t-1)}} p(\mathbf{X}_{n\bullet}^{(1:t-1)}, z_{nk}^{(t-1)}|-) p(z_{nk}^{(t)}|z_{nk}^{(t-1)}). \tag{8.80}$$

We perform a forward-filtering backward-sampling (FFBS) procedure in order to get samples from the posterior of the time-dependent matrix $\mathbf{Z}$.

Forward step. First, we compute the forward factors $f[z_{nk}^{(t)}] = p(z_{nk}^{(t)} = 1|\mathbf{X}_{n\bullet}^{(1:t)}, \mathbf{Z}_{n,\neg k}^{(t)}, \mathbf{B})$ and $(1 - f[z_{nk}^{(t)}]) = p(z_{nk}^{(t)} = 0|\mathbf{X}_{n\bullet}^{(1:t)}, \mathbf{Z}_{n,\neg k}^{(t)}, \mathbf{B})$ for each country n and timestep t:

$$p(z_{nk}^{(t)}|\mathbf{X}_{n\bullet}^{(1:t)}, \mathbf{Z}_{n,\neg k}^{(t)}, \mathbf{B}) \propto p(\mathbf{X}_{n\bullet}^{(t)}|\mathbf{Z}_{n\bullet}^{(t)}, \mathbf{B}) \cdot \sum_{z_{nk}^{(t-1)}} p(z_{nk}^{(t-1)}|\mathbf{X}_{n\bullet}^{(1:t-1)}, \mathbf{Z}_{n,\neg k}^{(t)}) p(z_{nk}^{(t)}|z_{nk}^{(t-1)}), \tag{8.81}$$

$$f[z_{nk}^{(t)}] = p(z_{nk}^{(t)} = 1|\mathbf{X}_{n\bullet}^{(1:t)}, \mathbf{Z}_{n,\neg k}^{(t)}, \mathbf{B}) = \frac{1}{1 + \mathrm{e}^{-u_{nk}^{(t)}}}, \tag{8.82}$$

$$\text{where } u_{nk}^{(t)} = \sum_{d=1}^{D} x_{nd}^{(t)} \log \left(\frac{\sum_{j \neq k} z_{nj}^{(t)} B_{jd} + B_{kd}}{\sum_{j \neq k} z_{nj}^{(t)} B_{jd}} \right) - \sum_{d=1}^{D} B_{kd} + \log \frac{a_k(1 - f[z_{nk}^{(t-1)}]) + b_k f[z_{nk}^{(t-1)}]}{(1 - a_k)(1 - f[z_{nk}^{(t-1)}]) + (1 - b_k) f[z_{nk}^{(t-1)}]}, \tag{8.83}$$

Backward step. Next, we iteratively sample from:

$$p(z_{nk}^{(t)}|z_{nk}^{(t+1)}, \mathbf{X}_{n\bullet}^{(1:t)}, \mathbf{Z}_{n,\neg k}^{(t)}, \mathbf{B}) \propto p(z_{nk}^{(t)}|\mathbf{X}_{n\bullet}^{(1:t)}, \mathbf{Z}_{n,\neg k}^{(t)}, \mathbf{B}) p(z_{nk}^{(t+1)}|z_{nk}^{(t)}) \tag{8.84}$$

$$p(z_{nk}^{(t)} = 1 | z_{nk}^{(t+1)}, \mathbf{X}_{n\bullet}^{(1:t)}, \mathbf{Z}_{n,\neg k}^{(t)}, \mathbf{B}) = \frac{1}{1 + \mathrm{e}^{-g_{nk}^{(t)}}}, \tag{8.85}$$

$$\text{where } g_{nk}^{(t)} = \log \frac{f[z_{nk}^{(t)}]}{1 - f[z_{nk}^{(t)}]} + z_{nk}^{(t+1)} \log \frac{b_k}{a_k} + (1 - z_{nk}^{(t+1)}) \log \frac{1 - b_k}{1 - a_k}, \tag{8.86}$$

Finally, we sample the activation probability a_k using a slice sampler, and persistence parameter b_k of each transition matrix:

$$\begin{aligned} p(a_k|-) &\propto p(\mathbf{Z}_{\bullet k}^{(\bullet)}|\mathbf{X}, \mu, \boldsymbol{a}, \boldsymbol{b})\; p(a_k) \\ &\propto \prod_{\forall n,t|z_{nk}^{(t-1)}=0}^{N} \prod^{T} p(z_{nk}^{(t)}|a_k)\; p(a_k) \\ &\propto \text{Beta}\left(\sum_{\forall n,t|z_{nk}^{(t-1)}=0}^{N} \sum^{T} z_{nk}^{(t)},\; 1 + m_k^0 \;-\; \sum_{\forall n,t|z_{nk}^{(t-1)}=0}^{N} \sum^{T} z_{nk}^{(t)} \right) \end{aligned} \tag{8.87}$$

$$\begin{aligned} p(b_k|-) &\propto p(\mathbf{Z}_{\bullet k}^{(\bullet)}|\mathbf{X}, \mu, \boldsymbol{a}, \boldsymbol{b})\; p(b_k) \\ &\propto \prod_{\forall n,t|z_{nk}^{(t-1)}=1}^{N} \prod^{T} p(z_{nk}^{(t)}|b_k)\; p(b_k) \\ &\propto \text{Beta}\left(\gamma + \sum_{\forall n,t|z_{nk}^{(t-1)}=1}^{N} \sum^{T} z_{nk}^{(t)},\; \delta + m_k^1 \;-\; \sum_{\forall n,t|z_{nk}^{(t-1)}=1}^{N} \sum^{T} z_{nk}^{(t)} \right) \end{aligned} \tag{8.88}$$

where $m_k^0 = \sum^N \sum^T \mathbb{1}\{z_{nk}^{(t-1)} = 0\}$ and $m_k^1 = \mathbb{1}\{\sum^N \sum^T z_{nk}^{(t-1)} = 1\}$.

The stick-breaking representation for the mIBP of [56] turns out to be very useful for inference purposes, as it allows us to use a combination of slice sampling and dynamic programming. The corresponding equations are the following:

$$a_1 \propto \text{Beta}(\alpha, 1) \tag{8.89}$$

$$p(a_k|a_{k-1}) = \alpha a_{k-1}^{-\alpha} a_k^{\alpha-1} \mathbb{1}(0 \leq a_k \leq a_{k-1}) \tag{8.90}$$

On the other hand, variables b_k are all independent of the number of latent features K, and are obtained as draws from a $\text{Beta}(\gamma, \delta)$ distribution. The central idea behind the slice sampler lies on

introducing an auxiliary slice variable μ with the following distribution:

$$\mu \sim \text{Uniform}(0, \min_{k:\exists t, \mathbf{Z}_{\bullet k}^{(t)}=1} a_k) \tag{8.91}$$

If we marginalize out the distribution $p(\mu, \boldsymbol{a}, \boldsymbol{b}, \mathbf{Z}) = p(\mu|\boldsymbol{a}, \mathbf{Z})p(\boldsymbol{a}, \boldsymbol{b}, \mathbf{Z})$ with respect to μ, it is clear that we recover the original joint distribution $p(\boldsymbol{a}, \boldsymbol{b}, \mathbf{Z})$. However, if we condition on μ, we get:

$$p(\mathbf{Z}|\mathbf{X}, \mu, \boldsymbol{a}, \boldsymbol{b}) \propto p(\mathbf{Z}|\mathbf{X}, \boldsymbol{a}, \boldsymbol{b}) \frac{\mathbb{1}(0 \leq \mu \leq \min_{k:\exists t, \mathbf{Z}_{\bullet k}^{(t)}=1} a_k)}{\min_{k:\exists t, \mathbf{Z}_{\bullet k}^{(t)}=1} a_k} \tag{8.92}$$

which forces all columns of $\mathbf{X}$ for which $a_k < \mu$ to be all zero. Note that there can only be a finite number of weights $a_k > \mu$, such that we only need to sample a finite (bounded) number of columns of $\mathbf{Z}$.

References

[1] Lauri A. Aaltonen, Paivi Peltomaki, Fredrick S. Leach, Pertti Sistonen, Lea Pylkkanen, Jukka-Pekka Mecklin, Heikki Jarvinen, Steven M. Powell, Jin Jen, Stanley R. Hamilton, et al., "Clues to the pathogenesis of familial colorectal cancer," *Science*, vol. 260, no. 5109, pp. 812–817, May 1993.

[2] Ghassan K. Abou-Alfa, Oscar Puig, Bruno Daniele, Masatoshi Kudo, Philippe Merle, Joong-Won Park, et al., "Randomized phase II placebo controlled study of Codrituzumab in previously treated patients with advanced hepatocellular carcinoma," *Journal of Hepatology*, Apr. 2016, doi:10.1016/j.jhep.2016.04.004.

[3] Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou, "Nonparametric Bayesian factor analysis for dynamic count matrices," *arXiv preprint*, Dec. 2015, arXiv:1512.08996.

[4] Tomasz Adamusiak and Mary Shimoyama, "EHR-based phenome wide association study in pancreatic cancer," *AMIA Summits on Translational Science Proceedings*, vol. 2014, pp. 9–15, Apr. 2014.

[5] Nibia Aires, "Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto sampling designs," *Methodology and Computing in Applied Probability*, vol. 1, no. 4, pp. 457–469, 1999, doi:10.1023/A:1010091628740.

[6] David J. Aldous, "Exchangeability and related topics," pp. 1–198, 1985, doi:10.1007/BFb0099421.

[7] Bernard W. E. Alford, *Britain in the World Economy since 1880*, Routledge, Nov. 1995, doi:10.4324/9781315837086.

[8] Mauricio A. Alvarez, David Luengo, Michalis K. Titsias, and Neil D. Lawrence, "Efficient multioutput Gaussian processes through variational inducing kernels.," *Artificial Intelligence and Statistics (AISTATS)*, vol. 9, pp. 25–32, 2010.

[9] R. Michael Alvarez and Jonathan Nagler, "Economics, issues and the Perot candidacy: voter choice in the 1992 presidential election," *American Journal of Political Science*, vol. 39, no. 3, pp. 714–744, 1995, doi:10.2307/2111651.

[10] Ulrika Andersson, Roberta McKean-Cowdin, Ulf Hjalmars, and Beatrice Malmer, "Genetic variants in association studies–review of strengths and weaknesses in study design and current

knowledge of impact on cancer risk," *Acta Oncologica (Stockholm, Sweden)*, vol. 48, no. 7, pp. 948–954, 2009.

[11] Elaine Angelino, Matthew James Johnson, and Ryan P. Adams, "Patterns of scalable Bayesian inference," Feb. 2016, arXiv:1602.05221.

[12] Charles E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, Nov. 1974, doi:10.1214/aos/1176342871.

[13] Cedric Archambeau, Balaji Lakshminarayanan, and Guillaume Bouchard, "Latent IBP compound Dirichlet allocation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 2, pp. 321–333, 2015.

[14] Alan R. Aronson, "Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program.," in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 17.

[15] Paramvir Bahl and Venkata N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," in *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, 2000, vol. 2, pp. 775–784, doi:10.1109/INFCOM.2000.832252.

[16] Bela Balassa, "The purchasing-power parity doctrine: A reappraisal," *Journal of Political Economy*, vol. 72, no. 6, pp. 584–596, 1964, doi:10.1086/258965.

[17] Yoav Benjamini and Yosef Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, Jan. 1995.

[18] Christopher M. Bishop, *Pattern Recognition*, vol. 128, 2006.

[19] David Blackwell and James B. MacQueen, "Ferguson distributions via Polya urn schemes," *The Annals of Statistics*, vol. 1, no. 2, pp. 353–355, Mar. 1973, doi:10.1214/aos/1176342372.

[20] David M. Blei, "Build, compute, critique, repeat: Data analysis with latent variable models," *Annual Review of Statistics and Its Application*, vol. 1, no. 1, pp. 203–232, 2014, doi:10.1146/annurev-statistics-022513-115657.

[21] David M. Blei and Michael I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian analysis*, vol. 1, no. 1, pp. 121–143, 2006.

[22] Olivier Bodenreider, "The Unified Medical Language System (UMLS): Integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. suppl_1, pp. D267–D270, Jan. 2004, doi:10.1093/nar/gkh061.

[23] Edwin V. Bonilla, Kian M. Chai, and Christopher Williams, "Multi-task Gaussian process prediction," in *Advances in Neural Information Processing Systems 20*, 2008, pp. 153–160.

[24] Henry J. Bruton, "Egypt's development in the seventies," *Economic Development and Cultural Change*, vol. 31, no. 4, pp. 679–704, 1983, doi:10.1086/451353.

[25] Cesar A. Calderon and Luis Servén, "Trends in infrastructure in latin america, 1980-2001," 2004, doi:https://doi.org/10.1596/1813-9450-3401.

[26] Mariana Capurro, Tonya Martin, Wen Shi, and Jorge Filmus, "Glypican-3 binds to Frizzled and plays a direct role in the stimulation of canonical Wnt signaling," *Journal of Cell Science*, vol. 127, no. Pt 7, pp. 1565–1575, Apr. 2014, doi:10.1242/jcs.140871.

[27] Carlos M. Carvalho, Jeffrey Chang, Joseph E. Lucas, Joseph R. Nevins, Quanli Wang, and Mike West, "High-dimensional sparse factor modeling: Applications in gene expression genomics," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1438–1456, Dec. 2008, doi:10.1198/016214508000000869.

[28] Mengjie Chen, Chao Gao, and Hongyu Zhao, "Phylogenetic Indian buffet process: Theory and applications in integrative analysis of cancer genomics," July 2013, arXiv:1307.8229.

[29] Donavan T. Cheng, Talia N. Mitchell, Ahmet Zehir, Ronak H. Shah, Ryma Benayed, Aijazuddin Syed, Raghu Chandramohan, Zhen Yu Liu, Helen H. Won, et al., "Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology," *The Journal of molecular diagnostics: JMD*, vol. 17, no. 3, pp. 251–264, May 2015, doi:10.1016/j.jmoldx.2014.12.006.

[30] Yeonseung Chung and David B. Dunson, "Nonparametric Bayes conditional distribution modeling with variable selection," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1646–1660, Dec. 2009, doi:10.1198/jasa.2009.tm08302.

[31] Geraldine M. Clarke, Carl A. Anderson, Fredrik H. Pettersson, Lon R. Cardon, Andrew P. Morris, and Krina T. Zondervan, "Basic statistical analysis in genetic case-control studies," *Nature protocols*, vol. 6, no. 2, pp. 121–133, Feb. 2011, doi:10.1038/nprot.2010.182.

[32] Francis S. Collins and Harold Varmus, "A new initiative on precision medicine," *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, Feb. 2015, doi:10.1056/NEJMp1500523.

[33] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sept. 1995, doi:10.1007/BF00994018.

[34] Andreas C. Damianou, Michalis K. Titsias, and Neil D. Lawrence, "Variational inference for latent variables and uncertain inputs in Gaussian processes," *Journal of Machine Learning Research*, vol. 17, no. 42, pp. 1–62, 2016.

[35] Bruno De Finetti, "Sur la condition d'equivalence partielle," *Trans. in Studies in Inductive Logic and Probability*, 1938.

[36] Aubrey D. N. J. de Grey, "Protagonistic pleiotropy: Why cancer may be the only pathogenic effect of accumulating nuclear mutations and epimutations in aging," *Mechanisms of Ageing and Development*, vol. 128, no. 7-8, pp. 456–459, Aug. 2007, doi:10.1016/j.mad.2007.05.005.

[37] Maria De Iorio, Peter Müller, Gary L. Rosner, and Steven N. MacEachern, "An ANOVA model for dependent random measures," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 205–215, 2004, doi:http://www.jstor.org/stable/27590366.

[38] Katrin Dippold and Harald Hruschka, "A parsimonious multivariate Poisson model for market basket analysis," *Review of Managerial Science*, vol. 7, no. 4, pp. 393–415, Aug. 2012, doi:10.1007/s11846-012-0088-7.

[39] Finale Doshi-Velez and Zoubin Ghahramani, "Accelerated sampling for the Indian buffet process," in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, ICML09, pp. 273–280, ACM, doi:10.1145/1553374.1553409.

[40] Finale Doshi-Velez and Zoubin Ghahramani, "Correlated non-parametric latent feature models," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 2009, pp. 143–150, AUAI Press.

[41] Finale Doshi-Velez and Been Kim, "Towards a rigorous science of interpretable machine learning," Feb. 2017, arXiv:1702.08608.

[42] Finale Doshi-Velez, Kurt Miller, Jurgen V. Gael, and Yee W. Teh, "Variational inference for the Indian buffet process," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 137–144.

[43] Finale Doshi-Velez and Sinead A. Williamson, "Restricted Indian buffet processes," Aug. 2015, arXiv:1508.06303.

[44] Jason A. Duan, Michele Guindani, and Alan E. Gelfand, "Generalized spatial Dirichlet process models," *Biometrika*, vol. 94, no. 4, pp. 809–825, Dec. 2007.

[45] Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth, "Hybrid Monte Carlo," *Physics letters B*, vol. 195, no. 2, pp. 216–222, 1987.

[46] David B. Dunson, "Nonparametric Bayes applications to biostatistics," *Bayesian nonparametrics*, vol. 28, pp. 223, 2010.

[47] David B. Dunson, Natesh Pillai, and Ju-Hyun Park, "Bayesian density regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 163–183, Apr. 2007, doi:10.1111/j.1467-9868.2007.00582.x.

[48] David K. Duvenaud, Hannes Nickisch, and Carl E. Rasmussen, "Additive Gaussian processes," in *Advances in Neural Information Processing Systems 24*, pp. 226–234. 2011.

[49] Douglas F. Easton and Rosalind A. Eeles, "Genome-wide association studies in cancer," *Human Molecular Genetics*, vol. 17, no. R2, pp. R109–R115, Oct. 2008, doi:10.1093/hmg/ddn287.

[50] Michael D. Escobar and Mike West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, pp. 577–588, 1994, doi:10.1.1.51.1747.

[51] Angela Fan, Finale Doshi-Velez, and Luke Miratrix, "Promoting domain-specific terms in topic models with informative priors," Jan. 2017, arXiv:1701.03227.

[52] Mohammad Farooq, Sun Young Hwang, Mi Kyung Park, Jung-Chul Kim, Moon Kyu Kim, and Young Kwan Sung, "Blocking endogenous glypican-3 expression releases Hep 3B cells from G1 arrest," *Molecules and Cells*, vol. 15, no. 3, pp. 356–360, June 2003.

[53] Thomas S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[54] Nicholas J. Foti and Sinead A. Williamson, "A survey of non-exchangeable priors for Bayesian nonparametric models," Nov. 2012, arXiv:1211.4798.

[55] Emily B. Fox and David B. Dunson, "Multiresolution Gaussian processes," Sept. 2012, arXiv:1209.0833.

[56] Jurgen V. Gael, Yee W. Teh, and Zoubin Ghahramani, "The infinite factorial hidden Markov model," in *Advances in Neural Information Processing Systems 21*, pp. 1697–1704. 2009.

[57] Jérôme Galon, Franck Pagès, Francesco M. Marincola, Magdalena Thurin, Giorgio Trinchieri, Bernard A. Fox, Thomas F. Gajewski, and Paolo A. Ascierto, "The immune score as a new possible approach for the classification of cancer," *Journal of Translational Medicine*, vol. 10, pp. 1, Jan. 2012, doi:10.1186/1479-5876-10-1.

[58] Alan E. Gelfand, Athanasios Kottas, and Steven N. MacEachern, "Bayesian nonparametric spatial modeling with Dirichlet process mixing," *Journal of the American Statistical Association*, vol. 100, no. 471, pp. 1021–1035, Sept. 2005.

[59] Alan E. Gelfand and Adrian F. M. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 398–409, 1990.

[60] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, *Bayesian Data Analysis*, vol. 2, CRC press Boca Raton, FL, 2014.

[61] Samuel J. Gershman and David M. Blei, "A tutorial on Bayesian nonparametric models," *Journal of Mathematical Psychology*, vol. 56, no. 1, pp. 1–12, 2012, doi:10.1016/j.jmp.2011.08.004.

[62] Zoubin Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015, doi:10.1038/nature14541.

[63] Zoubin Ghahramani and Carl E. Rasmussen, "Bayesian Monte Carlo," in *Advances in Neural Information Processing Systems 16*, 2003, pp. 489–496.

[64] Soumya Ghosh, Andrei B. Ungureanu, Erik B. Sudderth, and David M. Blei, "Spatial distance dependent Chinese restaurant processes for image segmentation," in *Advances in Neural Information Processing Systems 24*, pp. 1476–1484. 2011.

[65] Walter R. Gilks and Pascal Wild, "Adaptive rejection sampling for Gibbs sampling," *Applied Statistics*, vol. 41, no. 2, pp. 337, 1992, doi:10.2307/2347565.

[66] Bryce Goodman and Seth Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," June 2016, arXiv:1606.08813.

[67] Prem Gopalan, Laurent Charlin, and David M. Blei, "Content-based recommendations with Poisson factorization," in *Advances in Neural Information Processing Systems 27*, pp. 3176–3184. 2014.

[68] Prem Gopalan, Jake M. Hofman, and David M. Blei, "Scalable recommendation with Poisson factorization," 2013, arXiv:1311.1704.

[69] Prem Gopalan, Jake M. Hofman, and David M. Blei, "Scalable recommendation with hierarchical Poisson factorization," in *Proceedings of the Thirti-first Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-15). AUAI Press*, 2015.

[70] Prem Gopalan, Francisco J. R. Ruiz, Rajesh Ranganath, and David M. Blei, "Bayesian nonparametric Poisson factorization for recommendation systems," *Artificial Intelligence and Statistics (AISTATS)*, vol. 33, pp. 275–283, 2014.

[71] Jim E. Griffin and Mark F. J. Steel, "Bayesian nonparametric modelling with the Dirichlet process regression smoother," *Statistica Sinica*, vol. 20, no. 4, pp. 1507, 2010.

[72] Thomas L. Griffiths and Zoubin Ghahramani, "The Indian buffet process: An introduction and review.," *Journal of Machine Learning Research*, vol. 12, pp. 1185–1224, 2011.

[73] Peter D. Grünwald, In Jae Myung, and Mark A. Pitt, *Advances in Minimum Description Length: Theory and Applications*, MIT press, 2005.

[74] Arnaud Guille, Max Chaffanet, and Daniel Birnbaum, "Signaling pathway switch in breast cancer," *Cancer Cell International*, vol. 13, no. 1, pp. 66, 2013.

[75] Sandeep K. Gupta, "Use of Bayesian statistics in drug development: Advantages and challenges," *International Journal of Applied and Basic Medical Research*, vol. 2, no. 1, pp. 3–6, 2012, doi:10.4103/2229-516X.96789.

[76] Dilan Görür and Carl E. Rasmussen, "Dirichlet process Gaussian mixture models: Choice of the base distribution," *Journal of Computer Science and Technology*, vol. 25, no. 4, pp. 653–664, July 2010, doi:10.1007/s11390-010-9355-8.

[77] P. Le Hai-son and Ziv Bar-Joseph, "Inferring interaction networks using the IBP applied to micro-RNA target prediction," in *Advances in Neural Information Processing Systems 24*, 2011, pp. 235–243.

[78] Dorit Hammerling, Matthew Cefalu, Jessi Cisewski, Francesca Dominici, Giovanni Parmigiani, Charles Paulson, and Richard L. Smith, "Completing the results of the 2013 Boston marathon," *PLoS ONE*, vol. 9, no. 4, pp. e93800, Apr. 2014, doi:10.1371/journal.pone.0093800.

[79] Douglas Hanahan and Robert A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, Jan. 2000, doi:10.1016/S0092-8674(00)81683-9.

[80] Douglas Hanahan and Robert A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011, doi:10.1016/j.cell.2011.02.013.

[81] Yukihiro Haruyama, Kenji Yorita, Tetsuji Yamaguchi, Sachiko Kitajima, Jun Amano, Toshihiko Ohtomo, Akinobu Ohno, Kazuhiro Kondo, and Hiroaki Kataoka, "High preoperative levels of serum glypican-3 containing N-terminal subunit are associated with poor prognosis in patients with hepatocellular carcinoma after partial hepatectomy," *International Journal of Cancer*, vol. 137, no. 7, pp. 1643–1651, Oct. 2015, doi:10.1002/ijc.29518.

[82] Ricardo Hausmann and César A. Hidalgo, "The network structure of economic output," *Journal of Economic Growth*, vol. 16, no. 4, pp. 309–342, Oct. 2011, doi:10.1007/s10887-011-9071-4.

[83] Ricardo Hausmann, César A. Hidalgo, Sebastian Bustos, Michele Coscia, Sarah Chung, Juan Jimenez, Alexander Simoes, and Muhammed A. Yildirim, *The Atlas of Economic Complexity*, 2014.

[84] Ricardo Henao, James T. Lu, Joseph E. Lucas, Jeffrey Ferranti, and Lawrence Carin, "Electronic health record analysis via deep Poisson factor models," *Journal of Machine Learning Research*, 2015.

[85] N. Lynn Henry and Daniel F. Hayes, "Cancer biomarkers," *Molecular Oncology*, vol. 6, no. 2, pp. 140–146, Apr. 2012, doi:10.1016/j.molonc.2012.01.010.

[86] James Hensman, Magnus Rattray, and Neil D. Lawrence, "Fast nonparametric clustering of structured time-series," Jan. 2014, arXiv:1401.1605.

[87] César A. Hidalgo and Ricardo Hausmann, "The building blocks of economic complexity," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10570–10575, June 2009, doi:10.1073/pnas.0900943106.

[88] César A. Hidalgo, Bailey Klinger, A.-L. Barabási, and Ricardo Hausmann, "The product space conditions the development of nations," *Science*, vol. 317, no. 5837, pp. 482–487, July 2007, doi:10.1126/science.1144581.

[89] Joel N. Hirschhorn, Kirk Lohmueller, Edward Byrne, and Kurt Hirschhorn, "A comprehensive review of genetic association studies," *Genetics in Medicine*, vol. 4, no. 2, pp. 45–61, Mar. 2002, doi:10.1097/00125817-200203000-00002.

[90] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1303–1347, May 2013.

[91] Patrik O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Dec. 2004.

[92] Hui-Ling Huang, Yu-Chung Wu, Li-Jen Su, Yun-Ju Huang, Phasit Charoenkwan, Wen-Liang Chen, Hua-Chin Lee, William Cheng-Chung Chu, and Shinn-Ying Ho, "Discovery of prognostic biomarkers for predicting lung cancer metastasis using microarray and survival data," *BMC Bioinformatics*, vol. 16, no. 1, Feb. 2015, doi:10.1186/s12859-015-0463-x.

[93] Masafumi Ikeda, Shinichi Ohkawa, Takuji Okusaka, Shuichi Mitsunaga, Satoshi Kobayashi, Chigusa Morizane, Ikue Suzuki, Shunsuke Yamamoto, and Junji Furuse, "Japanese phase I study of GC33, a humanized antibody against glypican-3 for advanced hepatocellular carcinoma," *Cancer Science*, vol. 105, no. 4, pp. 455–462, Apr. 2014, doi:10.1111/cas.12368.

[94] Vesna Ilakovac, "Statistical hypothesis testing and some pitfalls," *Biochemia Medica*, vol. 19, no. 1, pp. 10–16, 2009, doi:10.1136/bmj.g5310.

[95] Yusuke Imamura, Shinichi Sakamoto, Takumi Endo, Takanobu Utsumi, Miki Fuse, Takahito Suyama, Koji Kawamura, Takashi Imamoto, Kojiro Yano, Katsuhiro Uzawa, et al., "FOXA1 promotes tumor progression in prostate cancer via the insulin-like growth factor binding protein 3 pathway," *PLoS ONE*, vol. 7, no. 8, pp. e42456, 2012, doi:10.1371/journal.pone.0042456.

[96] Ronald L. Iman and William J. Conover, "The use of the rank transform in regression," *Technometrics*, vol. 21, no. 4, pp. 499–509, 1979, doi:10.2307/1268289.

[97] Sumit Isharwal, François Audenet, Eugene J. Pietzak, Eugene K. Cha, Gopa Iyer, Ahmet Zehir, Barry S. Taylor, Michael F. Berger, Satish Tickoo, Victor E. Reuter, et al., "Comparison of genomic alterations in bladder urothelial tumors with and without telomerase reverse transcriptase promoter mutation using a next-generation sequencing assay," 2017.

[98] Takahiro Ishiguro, Masamichi Sugimoto, Yasuko Kinoshita, Yoko Miyazaki, Kiyotaka Nakano, Hiroyuki Tsunoda, Izumi Sugo, Iwao Ohizumi, Hiroyuki Aburatani, et al., "Anti-glypican 3 antibody as a potential antitumor agent for human liver cancer," *Cancer Research*, vol. 68, no. 23, pp. 9832–9838, Dec. 2008, doi:10.1158/0008-5472.CAN-08-1973.

[99] Hemant Ishwaran and Lancelot F. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, vol. 96, no. 453, 2001.

[100] J. Larry Jameson and Dan L. Longo, "Precision medicine: Personalized, problematic, and promising," *New England Journal of Medicine*, vol. 372, no. 23, pp. 2229–2234, June 2015, doi:10.1056/NEJMsb1503104.

[101] Peter B. Jensen, Lars J. Jensen, and Søren Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012, doi:10.1038/nrg3208.

[102] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, Nov. 1999, doi:10.1023/A:1007665907178.

[103] Peter Juni, Douglas G. Altman, and Matthias Egger, "Assessing the quality of controlled clinical trials," *British Medical Journal*, vol. 323, no. 7303, pp. 42, 2001, doi:10.1136/bmj.323.7303.42.

[104] Been Kim, Julie A Shah, and Finale Doshi-Velez, "Mind the gap: A generative approach to interpretable feature selection and extraction," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 2260–2268.

[105] John F. C. Kingman, "Completely random measures.," *Pacific Journal of Mathematics*, vol. 21, no. 1, pp. 59–78, 1967.

[106] John F. C. Kingman, *Poisson processes*, Wiley Online Library, 1993.

[107] David Knowles and Zoubin Ghahramani, "Nonparametric Bayesian sparse factor models with application to gene expression modeling," *The Annals of Applied Statistics*, vol. 5, no. 2B, pp. 1534–1552, June 2011, doi:10.1214/10-AOAS435.

[108] Vathany Kulasingam and Eleftherios P. Diamandis, "Strategies for discovering novel cancer biomarkers through utilization of emerging technologies," *Nature Clinical Practice Oncology*, vol. 5, no. 10, pp. 588–599, 2008, doi:10.1038/ncponc1187.

[109] Dean Lacy and Barry C. Burden, "The vote-stealing and turnout effects of Ross Perot in the 1992 U.S. presidential election," *American Journal of Political Science*, vol. 43, no. 1, pp. 233–255, 1999, doi:10.2307/2991792.

[110] Emilie Lalonde, Adrian S. Ishkanian, Jenna Sykes, Michael Fraser, Helen Ross-Adams, Nicholas Erho, Mark J. Dunning, Silvia Halim, Alastair D Lamb, et al., "Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: A retrospective cohort study," *The Lancet Oncology*, vol. 15, no. 13, pp. 1521–1532, Dec. 2014, doi:10.1016/S1470-2045(14)71021-6.

[111] Neil Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of Machine Learning Research*, vol. 6, no. Nov, pp. 1783–1816, 2005.

[112] Juhee Lee, Peter Müller, Kamalakar Gulukota, and Yuan Ji, "A Bayesian feature allocation model for tumor heterogeneity," *The Annals of Applied Statistics*, vol. 9, no. 2, pp. 621–639, June 2015, doi:10.1214/15-AOAS817.

[113] Jörg C. Lemm, "Mixtures of Gaussian process priors," Nov. 1999, arXiv:physics/9911077.

[114] Cathryn M. Lewis and Jo Knight, "Introduction to genetic association studies," *Cold Spring Harbor Protocols*, vol. 2012, no. 3, pp. 297–306, Mar. 2012, doi:10.1101/pdb.top068163.

[115] Peirce Lewis, Casey McCracken, and Roger Hunt, "Politics: Who cares," *American Demographics*, pp. 16–23, 1994.

[116] Madeline Li, Barbara Choo, Zeu-Ming Wong, Jorge Filmus, and Ronald N. Buick, "Expression of OCI-5/glypican 3 during intestinal morphogenesis: Regulation by cell shape in intestinal epithelial cells," *Experimental Cell Research*, vol. 235, no. 1, pp. 3–12, Aug. 1997, doi:10.1006/excr.1997.3629.

[117] Antonio Lijoi and Igor Prünster, "Models beyond the Dirichlet process," *Bayesian nonparametrics*, vol. 28, pp. 80, 2010.

[118] Dahua Lin, Eric Grimson, and John W. Fisher, "Construction of dependent Dirichlet processes based on Poisson processes," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 1396–1404.

[119] Fredrik Lindsten, Michael I. Jordan, and Thomas B. Schön, "Particle Gibbs with ancestor sampling.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2145–2184, 2014.

[120] Christoph Lippert, Francesco Paolo Casale, Barbara Rakitsch, and Oliver Stegle, "LIMIX: Genetic analysis of multiple traits," *bioRxiv*, p. 003905, May 2014, doi:10.1101/003905.

[121] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman, "FaST linear mixed models for genome-wide association studies," *Nature Methods*, vol. 8, no. 10, pp. 833–835, Oct. 2011, doi:10.1038/nmeth.1681.

[122] Kathryn L. Lunetta, "Genetic association studies," *Circulation*, vol. 118, no. 1, pp. 96–101, July 2008, doi:10.1161/CIRCULATIONAHA.107.700401.

[123] Miguel Lázaro-Gredilla, Steven Van Vaerenbergh, and Neil D. Lawrence, "Overlapping mixtures of Gaussian processes for the data association problem," *Pattern Recognition*, vol. 45, no. 4, pp. 1386–1395, Apr. 2012, doi:10.1016/j.patcog.2011.10.004.

[124] Steven N. MacEachern, "Dependent nonparametric processes," in *ASA proceedings of the section on Bayesian Statistical Science*, 1999, pp. 50–55.

[125] Steven N. MacEachern, "Dependent Dirichlet processes," *Unpublished manuscript, Department of Statistics, The Ohio State University*, 2000.

[126] Luca Martino, Jesse Read, and David Luengo, "Improved adaptive rejection metropolis sampling algorithms," *IEEE Transactions on Signal Processing*, vol. 63, no. 12, pp. 3123–3138, June 2015, doi:10.1109/TSP.2015.2420537.

[127] Marius Marusteri and Vladimir Bacarea, "Comparing groups for statistical differences: How to choose the right statistical test?," *Biochemia medica*, vol. 20, no. 1, pp. 15–32, 2010.

[128] Edward Meeds and Simon Osindero, "An alternative infinite mixture of Gaussian process experts," in *In Advances In Neural Information Processing Systems 19*, 2006.

[129] Stéphane M. Meystre, Guergana K. Savova, Karin C. Kipper-Schuler, John F. Hurdle, and others, "Extracting information from textual documents in the electronic health record: A review of recent research," *Yearb Med Inform*, vol. 35, pp. 128–44, 2008.

[130] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang, "Universal kernels," *Journal of Machine Learning Research*, vol. 7, pp. 2651–2667, 2006.

[131] Kurt T. Miller, Thomas Griffiths, and Michael I. Jordan, "The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features," *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI2008)*, 2012, arXiv:1206.3279.

[132] Kurt T. Miller, Michael I. Jordan, and Thomas L. Griffiths, "Nonparametric latent feature models for link prediction," in *Advances in Neural Information Processing Systems 22*, pp. 1276–1284. Curran Associates, Inc., 2009.

[133] Veena Mishra, "Indonesia: adjustment in the 1980s," in *Economic Restructuring in East Asia and India*, pp. 103–133. Palgrave Macmillan UK, 1995, doi:10.1057/9780230376038$_4$.

[134] Andriy Mnih and Ruslan R. Salakhutdinov, "Probabilistic matrix factorization," in *Advances in Neural Information Processing Systems 21*, 2008, pp. 1257–1264.

[135] Kevin P. Murphy, *Machine learning: A probabilistic perspective*, MIT press, 2012.

[136] Meera Nair, Sardul Singh Sandhu, and Anil K. Sharma, "Prognostic and predictive biomarkers in cancer," *Current Cancer Drug Targets*, vol. 14, no. 5, pp. 477–504, 2014.

[137] Kiyotaka Nakano, Tetsuro Orita, Junichi Nezu, Takeshi Yoshino, Iwao Ohizumi, Masamichi Sugimoto, Koh Furugaki, Yasuko Kinoshita, Takahiro Ishiguro, et al., "Anti-glypican 3 antibodies cause ADCC against human hepatocellular carcinoma cells," *Biochemical and Biophysical Research Communications*, vol. 378, no. 2, pp. 279–284, Jan. 2009, doi:10.1016/j.bbrc.2008.11.033.

[138] Tetsuya Nakatsura and Yasuharu Nishimura, "Usefulness of the novel oncofetal antigen glypican-3 for diagnosis of hepatocellular carcinoma and melanoma," *BioDrugs: Clinical Immunotherapeutics, Biopharmaceuticals and Gene Therapy*, vol. 19, no. 2, pp. 71–77, 2005.

[139] Ewelina Nalejska, Ewa Maczynska, and Marzena Anna Lewandowska, "Prognostic and predictive biomarkers: Tools in personalized oncology," *Molecular Diagnosis & Therapy*, vol. 18, no. 3, pp. 273–284, 2014, doi:10.1007/s40291-013-0077-9.

[140] Radford M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[141] Michael L. Nickerson, Garrett M. Dancik, Kate M. Im, Michael G. Edwards, Sevilay Turan, Joseph Brown, Christina Ruiz-Rodriguez, Charles Owens, James C. Costello, Guangwu Guo, et al., "Concurrent alterations in TERT, KDM6A, and the BRCA pathway in bladder cancer," *Clinical Cancer Research*, vol. 20, no. 18, pp. 4935–4948, 2014.

[142] Magali Olivier, Monica Hollstein, and Pierre Hainaut, "TP53 mutations in human cancers: Origins, consequences, and clinical use," *Cold Spring Harbor perspectives in biology*, vol. 2, no. 1, pp. a001008, 2010, doi:10.1101/cshperspect.a001008.

[143] Peter Orbanz and Yee Whye Teh, "Bayesian nonparametric models," in *Encyclopedia of Machine Learning*, pp. 81–89. Springer, 2010.

[144] Edoardo Otranto and Giampiero M. Gallo, "A nonparametric Bayesian approach to detect the number of regimes in Markov switching models," *Econometric Reviews*, vol. 21, no. 4, pp. 477–496, 2002.

[145] John W. Paisley, David M. Blei, and Michael I. Jordan, "Stick-breaking Beta processes and the Poisson process," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 850–858.

[146] Gaurav Pandey and Ambedkar Dukkipati, "On collapsed representation of hierarchical completely random measures," Sept. 2015, arXiv:1509.01817.

[147] Leopold Parts, Oliver Stegle, John Winn, and Richard Durbin, "Joint genetic analysis of gene expression data with inferred cellular phenotypes," *PLoS Genetics*, vol. 7, no. 1, pp. e1001276, Jan. 2011, doi:10.1371/journal.pgen.1001276.

[148] Jai N. Patel, Howard L. McLeod, and Federico Innocenti, "Implications of genome-wide association studies in cancer therapeutics," *British Journal of Clinical Pharmacology*, vol. 76, no. 3, pp. 370–380, Sept. 2013, doi:10.1111/bcp.12166.

[149] Sarah Payton, "Bladder cancer: Mutation found in $>$ 70% of tumours," *Nature Reviews Urology*, vol. 10, no. 11, pp. 616–616, 2013, doi:10.1038/nrurol.2013.222.

[150] Thomas A. Pearson and Teri A. Manolio, "How to interpret a genome-wide association study," *JAMA*, vol. 299, no. 11, pp. 1335–1344, Mar. 2008, doi:10.1001/jama.299.11.1335.

[151] Paul D. P. Pharoah, Alison M. Dunning, Bruce A. J. Ponder, and Douglas F. Easton, "Association studies for finding cancer-susceptibility genetic variants," *Nature Reviews Cancer*, vol. 4, no. 11, pp. 850–860, Nov. 2004, doi:10.1038/nrc1476.

[152] Jim Pitman, "Poisson–Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition," *Combinatorics, Probability and Computing*, vol. 11, no. 5, pp. 501–514, Sept. 2002, doi:10.1017/S0963548302005163.

[153] Jim Pitman, *Combinatorial stochastic processes: école d'été de probabilités de Saint-Flour XXXII-2002*, Springer, 2006.

[154] Melanie F. Pradier, Theofanis Karaletsos, Stefan Stark, Julia E. Vogt, Gunnar Rätsch, and Fernando Perez-Cruz, "Bayesian Poisson factorization for genetic associations with clinical features in cancer," in *Machine Learning for Healthcare Workshop in Neural Information Processing Systems*, 2015.

[155] Melanie F. Pradier, Pablo G. Moreno, Francisco J. R. Ruiz, Isabel Valera, Harold Molina-Bulla, and Fernando Perez-Cruz, "Map/reduce uncollapsed Gibbs sampling for Bayesian nonparametric models," in *Software Engineering for Machine Learning Workshop in Neural Information Processing Systems*, 2014.

[156] Melanie F. Pradier, Pablo M. Olmos, and Fernando Perez-Cruz, "Entropy-constrained scalar quantization with a lossy-compressed bit," *Entropy*, vol. 18, no. 12, pp. 449, 2016, doi:10.3390/e18120449.

[157] Melanie F. Pradier and Fernando Perez-Cruz, "Infinite mixture of global Gaussian processes," in *Bayesian Non-parametric: the Next Generation Workshop in Neural Information Processing Systems*, 2015.

[158] Melanie F. Pradier, Bernhard Reis, Lori Jukofsky, Francesca Milletti, Toshihiko Ohtomo, Fernando Perez-Cruz, and Oscar Puig, "Indian Buffet process identifies NK cell biomarkers as predictors of response to Codrituzumab in patients with advanced hepatocellular carcinoma.," *Submitted to BMC Cancer*, July 2017.

[159] Melanie F. Pradier, Francisco J. R. Ruiz, and Fernando Perez-Cruz, "Prior design for dependent Dirichlet processes: An application to marathon modeling," *PLoS ONE*, vol. 11, no. 1, pp. e0147402, Jan. 2016, doi:10.1371/journal.pone.0147402.

[160] Melanie F. Pradier, Stefan Stark, Stephanie Hyland, Julia E. Vogt, and Gunnar Rätsch, "Large-scale sentence clustering from electronic health records for genetic associations in cancer," in *Machine Learning for Computational Biology Workshop in Neural Information Processing Systems*, 2015.

[161] Melanie F. Pradier, Viktor Stojkoski, Zoran Utkovski, Lujupco Kocarev, and Fernando Perez-Cruz, "Sparse three-parameter restricted Indian buffet process for understanding international trade," in *Submitted to Advances in Neural Information Processing Systems*, 2017.

[162] Shaan Qamar and Surya T. Tokdar, "Additive Gaussian process regression," Nov. 2014, arXiv:1411.7009.

[163] Novi Quadrianto, Viktoriia Sharmanska, David A. Knowles, and Zoubin Ghahramani, "The supervised IBP: Neighbourhood preserving infinite latent feature models," Sept. 2013, arXiv:1309.6858.

[164] Elsa Quintana, Mark Shackleton, Hannah R. Foster, Douglas R. Fullen, Michael S. Sabel, Timothy M. Johnson, and Sean J. Morrison, "Phenotypic heterogeneity among tumorigenic melanoma cells from patients that is reversible and not hierarchically organized," *Cancer Cell*, vol. 18, no. 5, pp. 510–523, Nov. 2010, doi:10.1016/j.ccr.2010.10.012.

[165] Joaquin Quiñonero-Candela and Carl E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *The Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[166] Wullianallur Raghupathi and Viju Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Information Science and Systems*, vol. 2, pp. 3, 2014, doi:10.1186/2047-2501-2-3.

[167] Yohan Suryo Rahmanto, Jin-Gyoung Jung, Ren-Chin Wu, Yusuke Kobayashi, Christopher M. Heaphy, Alan K. Meeker, Tian-Li Wang, and Ie-Ming Shih, "Inactivating ARID1A tumor suppressor enhances TERT transcription and maintains telomere length in cancer cells," *Journal of Biological Chemistry*, vol. 291, no. 18, pp. 9690–9699, 2016.

[168] Rajesh Ranganath and David M. Blei, "Correlated random measures," July 2015, arXiv:1507.00720.

[169] Rajesh Ranganath, Sean Gerrish, and David M. Blei, "Black box variational inference," in *Artificial Intelligence and Statistics*, 2014, pp. 814–822, arXiv:1401.0118.

[170] Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David Blei, "Deep exponential families," in *Artificial Intelligence and Statistics*, 2015, pp. 762–771, arXiv:1411.2581.

[171] Carl E. Rasmussen and Zoubin Ghahramani, "Infinite mixtures of Gaussian process experts," *Advances in Neural Information Processing Systems 15*, vol. 2, pp. 881–888, 2002.

[172] Carl E. Rasmussen and Christopher K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2005.

[173] Priyadip Ray, Lingling Zheng, Joseph Lucas, and Lawrence Carin, "Bayesian joint analysis of heterogeneous genomics data," *Bioinformatics*, p. btu064, Jan. 2014, doi:10.1093/bioinformatics/btu064.

[174] Lu Ren, David B. Dunson, Scott Lindroth, and Lawrence Carin, "Dynamic nonparametric Bayesian models for analysis of music," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 458–472, 2010.

[175] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016, KDD '16, pp. 1135–1144, ACM, doi:10.1145/2939672.2939778.

[176] Marylyn D. Ritchie, Lance W. Hahn, Nady Roodi, L. Renee Bailey, William D. Dupont, Fritz F. Parl, and Jason H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *American Journal of Human Genetics*, vol. 69, no. 1, pp. 138–147, July 2001, doi:10.1086/321276.

[177] Christian P. Robert and George Casella, *Monte Carlo Statistical Methods (Springer Texts in Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[178] Abel Rodríguez, David B. Dunson, and Alan E. Gelfand, "Bayesian nonparametric functional data analysis through density estimation," *Biometrika*, vol. 96, no. 1, pp. 149–162, Mar. 2009.

[179] William F. Rosenberger and John M. Lachin, *Randomization in Clinical Trials: Theory and Practice*, John Wiley & Sons, Nov. 2015.

[180] James Ross and Jennifer Dy, "Nonparametric mixture of Gaussian processes with constraints," *JMLR Workshop and Conference Proceedings*, vol. 28, pp. 1346–1354, 2013.

[181] Sam T. Roweis and Lawrence K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000, doi:10.1126/science.290.5500.2323.

[182] Francisco J. R. Ruiz, Isabel Valera, Carlos Blanco, and Fernando Perez-Cruz, "Bayesian nonparametric comorbidity analysis of psychiatric disorders," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1215–1247, Jan. 2014.

[183] Lori C. Sakoda, Eric Jorgenson, and John S. Witte, "Turning of COGS moves forward findings for hormonally mediated cancers," *Nature Genetics*, vol. 45, no. 4, pp. 345–348, Apr. 2013, doi:10.1038/ng.2587.

[184] Aaron Schein, Hanna Wallach, and Mingyuan Zhou, "Poisson-Gamma dynamical systems," in *Advances in Neural Information Processing Systems 29*, pp. 5006–5014. 2016.

[185] Mikkel N. Schmidt, Ole Winther, and Lars Kai Hansen, "Bayesian non-negative matrix factorization.," in *ICA*. 2009, vol. 9, pp. 540–547, Springer.

[186] R. Schulz and C. Curnow, "Peak performance and age among superathletes: track and field, swimming, baseball, tennis, and golf," *Journal of Gerontology*, vol. 43, no. 5, pp. P113–120, Sept. 1988.

[187] Steven L. Scott, "Bayesian methods for hidden Markov models: Recursive computing in the 21st century," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 337–351, 2002.

[188] Jayaram Sethuraman, "A constructive definition of Dirichlet priors," *Statistica sinica*, vol. 4, no. 2, pp. 639–650, 1994.

[189] Amar Shah and Zoubin Ghahramani, "Markov Beta processes for time evolving dictionary learning," *Proceedings of Uncertainty in Artificial Inteligence*, 2016.

[190] Jian Qing Shi, Roderick Murray-Smith, and D. M. Titterington, "Hierarchical Gaussian process mixtures for regression," *Statistics and Computing*, vol. 15, no. 1, pp. 31–41, Jan. 2005, doi:10.1007/s11222-005-4787-7.

[191] David J. Spiegelhalter, Keith R. Abrams, and Jonathan P. Myles, *Bayesian approaches to clinical trials and health-care evaluation*, vol. 13, John Wiley & Sons, 2004.

[192] Viktor Stojkoski, Zoran Utkovski, and Ljupco Kocarev, "The impact of services on economic complexity: Service sophistication as route for economic growth," 2016, arXiv:1604.06284.

[193] Tom Strachan and Andrew Read, *Human Molecular Genetics*, New York: Garland Science/Taylor & Francis Group, c2011, United States, 4th ed. edition, 2010.

[194] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–724, Apr. 2009, doi:10.1038/nature07943.

[195] Shiliang Sun and Xin Xu, "Variational inference for infinite mixtures of Gaussian processes with applications to traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 466–475, June 2011, doi:10.1109/TITS.2010.2093575.

[196] Andrea Tacchella, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, and Luciano Pietronero, "A new metrics for countries' fitness and products' complexity," *Scientific Reports*, vol. 2, Oct. 2012, doi:10.1038/srep00723.

[197] Hirotake Takai, Motohki Ashihara, Takahiro Ishiguro, Hiromichi Terashima, Takeshi Watanabe, Atsuhiko Kato, and Masami Suzuki, "Involvement of glypican-3 in the recruitment of M2-polarized tumor-associated macrophages in hepatocellular carcinoma," *Cancer Biology & Therapy*, vol. 8, no. 24, pp. 2329–2338, Dec. 2009.

[198] Hirotake Takai, Atsuhiko Kato, Yasuko Kinoshita, Takahiro Ishiguro, Yayoi Takai, Yoshimi Ohtani, Masamichi Sugimoto, and Masami Suzuki, "Histopathological analyses of the antitumor activity of anti-glypican-3 antibody (GC33) in human liver cancer xenograft models: The contribution of macrophages," *Cancer Biology & Therapy*, vol. 8, no. 10, pp. 930–938, May 2009.

[199] Aditya Tayal, Pascal Poupart, and Yuying Li, "Hierarchical double Dirichlet process mixture of Gaussian processes.," in *AAAI*, 2012.

[200] Yee Whye Teh, "Dirichlet process," in *Encyclopedia of Machine Learning*, pp. 280–287. Springer US, 2011.

[201] Yee Whye Teh and Dilan Gorur, "Indian buffet processes with power-law behavior," in *Advances in Neural Information Processing Systems 22*, pp. 1838–1846. 2009.

[202] Yee Whye Teh, Dilan Gorur, and Zoubin Ghahramani, "Stick-breaking construction for the Indian buffet process," in *International Conference on Artificial Intelligence and Statistics*, 2007, pp. 556–563.

[203] Yee Whye Teh and Michael I. Jordan, *Bayesian Nonparametrics*, Cambridge University Press, 2010.

[204] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006, doi:10.1198/016214506000000302.

[205] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000, doi:10.1126/science.290.5500.2319.

[206] Romain Thibaux and Michael I. Jordan, "Hierarchical Beta processes and the Indian buffet process," in *International Conference on Artificial Intelligence and Statistics*, 2007, pp. 564–571.

[207] Michalis K. Titsias, "The infinite Gamma-Poisson feature model," in *Advances in Neural Information Processing Systems 20*, 2007, pp. 1513–1520.

[208] Isabel Valera, Melanie F. Pradier, and Zoubin Ghahramani, "General latent feature modeling for data exploration tasks," *Workshop on Human Iinterpretability in Machine Learning at Neural Information Processing Systems*, 2017, arXiv:1707.08352.

[209] Isabel Valera and Zoubin Ghahramani, "General table completion using a Bayesian nonparametric model," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 981–989.

[210] Isabel Valera, Melanie F. Pradier, and Zoubin Ghaharamani, "General latent feature model for heterogeneous datasets," *Submitted to Journal of Machine Learning Research*, 2017, arXiv:1706.03779.

[211] Isabel Valera, Francisco J. R. Ruiz, Pablo M. Olmos, Carlos Blanco, and Fernando Perez-Cruz, "Infinite continuous feature model for psychiatric comorbidity analysis," *Neural Computation*, pp. 1–28, Dec. 2015, doi:10.1162/NECO$_{a0}$0805.

[212] Isabel Valera, Francisco J. R. Ruiz, Lennart Svensson, and Fernando Perez-Cruz, "Infinite factorial dynamical model," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 1666–1674.

[213] Xiaoyue Wang, Audrey Q. Fu, Megan E. McNerney, and Kevin P. White, "Widespread genetic epistasis among cancer genes," *Nature Communications*, vol. 5, pp. 4828, Nov. 2014, doi:10.1038/ncomms5828.

[214] Zhuoran Wang, Anoop D. Shah, A. Rosemary Tate, Spiros Denaxas, John Shawe-Taylor, and Harry Hemingway, "Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning," *PLoS ONE*, vol. 7, no. 1, pp. e30412, Jan. 2012, doi:10.1371/journal.pone.0030412.

[215] Larry Wasserman, *All of Nonparametric Statistics, ser*, Springer Texts in Statistics. New York: Springer-Verlag, 2006.

[216] Larry Wasserman, *All of statistics: A concise course in statistical inference*, Springer Science & Business Media, 2013.

[217] Max Welling and Yee Whye Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 681–688.

[218] Grant R. Wilkinson, "Drug metabolism and variability among patients in drug response," *New England Journal of Medicine*, vol. 352, no. 21, pp. 2211–2221, May 2005, doi:10.1056/NEJMra032424.

[219] Sinead A. Williamson, Peter Orbanz, and Zoubin Ghahramani, "Dependent Indian buffet processes," in *International conference on artificial intelligence and statistics*, 2010, pp. 924–931.

[220] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S. Huang, and Shuicheng Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, June 2010, doi:10.1109/JPROC.2010.2044470.

[221] Naoko Yamauchi, Akira Watanabe, Michiyo Hishinuma, Ken-Ichi Ohashi, Yutaka Midorikawa, Yasuyuki Morishita, Toshiro Niki, Junji Shibahara, Masaya Mori, et al., "The glypican 3 oncofetal protein is a promising diagnostic marker for hepatocellular carcinoma," *Modern Pathology*, vol. 18, no. 12, pp. 1591–1598, Dec. 2005, doi:10.1038/modpathol.3800436.

[222] Christopher Yau, Omiros Papaspiliopoulos, Gareth O. Roberts, and Christopher Holmes, "Bayesian non-parametric hidden Markov models with applications in genomics," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 1, pp. 37–57, Jan. 2011, doi:10.1111/j.1467-9868.2010.00756.x.

[223] I-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," *Cement and Concrete Research*, vol. 28, no. 12, pp. 1797–1808, Dec. 1998, doi:10.1016/S0008-8846(98)00165-3.

[224] Wei Zhang, Jun Zhu, Eric E. Schadt, and Jun S. Liu, "A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules," *PLoS Computational Biology*, vol. 6, no. 1, Jan. 2010, doi:10.1371/journal.pcbi.1000642.

[225] Yu Zhang and Jun S Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nature Genetics*, vol. 39, no. 9, pp. 1167–1173, Sept. 2007, doi:10.1038/ng2110.

[226] Mingyuan Zhou, Lauren Hannah, David B. Dunson, and Lawrence Carin, "Beta-negative binomial process and Poisson factor analysis," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 1462–1471, arXiv:1112.3605.

[227] Andrew X. Zhu, Philip J. Gold, Anthony B. El-Khoueiry, Thomas A. Abrams, Hideo Morikawa, Norihisa Ohishi, Toshihiko Ohtomo, and Philip A. Philip, "First-in-man phase I study of GC33, a novel recombinant humanized antibody against glypican-3, in patients with advanced hepatocellular carcinoma," *Clinical Cancer Research*, vol. 19, no. 4, pp. 920–928, Feb. 2013, doi:10.1158/1078-0432.CCR-12-2616.