LÉVY RANDOM MEASURES: POSTERIOR CONSISTENCY AND APPLICATIONS

by

Natesh S. Pillai

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Dr. Robert L. Wolpert, Supervisor

_____
Dr. Merlise A. Clyde

_____
Dr. Sayan Mukherjee

_____
Dr. Subhashis Ghosal

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2008

ABSTRACT

(Statistics)

LÉVY RANDOM MEASURES: POSTERIOR CONSISTENCY AND APPLICATIONS

by

Natesh S. Pillai

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Dr. Robert L. Wolpert, Supervisor

_____
Dr. Merlise A. Clyde

_____
Dr. Sayan Mukherjee

_____
Dr. Subhashis Ghosal

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Department of Statistical Science in the Graduate School of
Duke University

2008

Abstract

Non-parametric function estimation using Lévy random measures is a very active area of current research. In this thesis further contributions, both theoretical and methodological, are made towards non-parametric function estimation using Lévy random measures.

In chapter 2, it is observed that Lévy random measures lead to a unified perspective of non-parametric function estimation using Bayesian methods and those using kernel methods such as Tikhonov regularization used in the machine learning literature. A coherent Bayesian kernel model based on an integral operator defined as the convolution of a kernel with a signed measure is studied. A few results on Fredholm integral operators are derived and a general class of measures whose image is dense in the reproducing kernel Hilbert space (RKHS) induced by the kernel is identified. These results lead to a function theoretic foundation for using non-parametric prior specifications in Bayesian modeling, such as Gaussian process and Dirichlet process prior distributions.

In chapter 3, easily verifiable conditions are derived for posterior consistency to hold in commonly used regression models with prior distributions on infinite dimensional spaces constructed from Lévy random fields. On route to proving conistency, convergence properties of finite dimensional approximations of Lévy random fields are studied. The key technical issues involved are outlined, and the results are illustrated by proving the posterior consisteny in concrete examples.

In chapter 4, the posterior consistency for non-parametric Poisson regression models is proved. The key step is to construct test functions that separate points, and have exponentially decaying type I and II errors.

In chapter 5, a novel application of Lévy random measures is discussed. It is

shown that Lévy random measures can be used for constructing prior distributions for spectral measures. Together with Bochner's theorem, this leads to a construction of non-parametric prior distributions on the cone of positive definite functions.

Contents

List of Figures

Acknowledgements

First and foremost, I would like to thank my advisor Dr. Robert L. Wolpert for patiently teaching me so many things over the past 5 years. I am very fortunate to get an advisor who is both a terrific mathematician and an equally brilliant statistician which gave me the confidence and perspective to work on research problems which were both technically challenging and also very important from a statistical view point. His attitude and enthusiasm towards research are inspring and I hope to continue learning from him. Thank you, Robert!

I thank my Ph.D. committee: Dr. Merlise A. Clyde for always being availabe for questions and discussing ideas; Dr Sayan Mukherjee for his cheerful spirit, many discussions on mathematics and for many other things; Dr. Subhashis Ghosal for his extremely useful comments which helped me improve this thesis tremendously.

A special thanks go to Dr. Michael J. Steele who first introduced me to probability and inspired me to go for a Ph.D. in statistics! The few weeks I had worked with him are a major highlight of my academic life so far. Dr. Jayanta Ghosh has been an inspiration for me to work on Bayesian non-parametrics and has always pointed me in the right research directions.

I have been fortunate to a few wonderful collaborators/mentors at Duke. Dr. Jonathan Mattingly has never ceased to amaze me with his ability to get to the heart of a problem quickly and make it look so easy! Dr. David Dunson has taught me many interesting things in Bayesian non-parametrics. I also would like to thank David for supervising my first published article. Dr. Alan Gelfand has always been a source of inspiration. I would also like to thank Dr. Michael Lavine, Dr. Mark Huber, Dr. Andrew Stuart and Dr. Ian Dinwoodie for many educative discussions. Dr Qiang Wu and Dr. Scott Mckinely has been so patient and put up with many

of my mathematical rants. Duke statistics and mathematics have wonderful faculty and I thank them all for being so helpful and making graduate school so much fun. The staff at Duke statistics department also made my life so much easier.

I have also made a few good friends at Duke. Abel Rodriguez always inspires me to work hard. I will miss both the statistical discussions and the lighter moments I had with Gavino Puggioni. Scotland Leman and Dawn Woodard has inspired me in more ways than they will realize. Eric Laber made my otherwise dull third-year so much fun when we shared an office and continues to amaze me with his work ethic. I also would like to thank my colleagues and office mates (Simon, Hongxia, Scott, Hao, Avishek, Chunlin) for their support and encouragement.

There are three special people who really enriched my life: Dr. Dane Claussen is an energizer bunny, a serious scholar and a best friend all morphed into one. Dr. Michael Hatch always makes me smile with his remarkable wit, competitive spirit and enthusiasm. Joyee Ghosh personifies a true friend, humility and kindness.

My undergraduate friends: Sriram, Aswin, Vivek, Ashutosh, Avartan, Sunil, Karthik , Nithin, Siddharth, Prakash and JamVwing group, deserve a special mention here. You have kept me going in crucial situations! My college friends Reghu and Arvind remain supportive as ever.

I would like to thank my Dad (Sivasubramonia Pillai) & Mom (Girija Kumari) for being so unconditionally supportive and loving, for their sacrifices to ensure that I received a first class education, and for always being there for me. Dad had always inspired me with his hard work. I also would like to thank my brother-in-law (Vijaykumar) and sister (Sreeja), cousins and the family back home for being so supportive in my endeavours. Finally, I dedicate this thesis to my ever-loving maternal grand parents - Shanmugham Pillai & Gomathy Ammal.

Chapter 1

Introduction

Non-parametric function estimation using Lévy random measures is a very active area of current research. Lévy random measures, derived from the classic infinitely divisible processes, are used for constructing prior distributions on function spaces based on overcomplete dictionaries. Overcomplete representations lead to possibly non-unique coefficients, but also facilitate sparser representations by using fewer non-zero coefficients. In most cases closed form expressions for posterior distributions are not available, but the prior construction using Lévy random fields permits tractable posterior simulation via a reversible jump Markov chain Monte Carlo algorithm. Powerful algorithms for efficient computation are available, and Lévy random field priors perform very well in comparison with other competing methods (Clyde and Wolpert [2007]).

1.1 Thesis outline

In this thesis further contributions, both theoretical and methodological, are made toward non-parametric function estimation using Lévy random measures.

In chapter 2, it is observed that Lévy random measures lead to a unified perspective of non-parametric function estimation using Bayesian methods and those using

kernel methods such as Tikhonov regularization commonly used in the machine learning literature. In this chapter, a coherent Bayesian kernel model based on an integral operator defined as the convolution of a kernel with a signed measure is studied. A general class of measures whose image is dense in the reproducing kernel Hilbert space (RKHS) induced by the kernel is identified. A consequence of this result is a function theoretic foundation for using non-parametric prior specifications in Bayesian modeling, such as Gaussian process and Dirichlet process prior distributions.

Although the Lévy random field priors work well in pratice, their posterior consistency has not been studied before in the literature. In chapter 3, easily verifiable conditions are derived for posterior consistency to hold in commonly used regression models. On route to proving consistency, the convergence properties of finite dimensional approximations of Lévy random fields are studied. The results are illustrated by proving the posterior consistency of two regression models with Lévy random field priors.

In chapter 4, the posterior consistency of a non-parametric Poisson regression model is proved. The key step is separating points in the parameter space by constructing hypothesis tests with suitably small error rates; this is done for individual pairs of points, and a covering argument is then used to show that the tests have exponentially decaying errors of types I and II.

In chapter 5, a novel application of Lévy random measures is discussed. It is shown that Lévy random measures can be used for constructing prior distributions for spectral measures. Together with Bochner's theorem, this leads to a construction of non-parametric prior distributions on the cone of positive definite functions. Some examples are discussed, and the methodology is illustrated by estimating the covariance function of one dimensional stationary Gaussian processes.

Chapter 2

Characterizing The Function Space For Bayesian Kernel Models

2.1 Introduction

Kernel methods have a long history in statistics and applied mathematics [Schoenberg 1942, Aronszajn 1950, Parzen 1963, de Boor and Lynch 1966, Micchelli and Wahba 1981, Wahba 1990] and have had a tremendous resurgence in the machine learning literature in the last ten years [Poggio and Girosi 1990, Vapnik 1998, Schölkopf and Smola 2001, Shawe-Taylor and Cristianini 2004]. Much of this resurgence was due to the popularization of classification algorithms such as support vector machines (SVMs) [Cortes and Vapnik 1995] that are particular instantiations of the method of regularization of Tikhonov [1963]. Many machine learning algorithms and statistical estimators can be summarized by the following penalized loss functional [Evgeniou *et al.* 2000, Hastie *et al.* 2001, Section 5.8]

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \left[L(f, \text{data}) + \lambda \|f\|_K^2 \right], \tag{2.1}$$

where L is a loss function, $\mathcal{H}$ is often an infinite-dimensional reproducing kernel Hilbert space (RKHS), $\|f\|_K^2$ is the norm of a function in this space, and λ is a tuning

parameter chosen to balance the trade-off between fitting errors and the smoothness of the function. The data is assumed to be drawn independently from a distribution $\rho(x, y)$ with $x \in \mathcal{X} \subset \mathbb{R}^d$ and $y \in \mathcal{Y} \subset \mathbb{R}$. Due to the representer theorem [Kimeldorf and Wahba 1971] the solution of the penalized loss functional will be a kernel

$$\hat{f}(x) = \sum_{i=1}^{n} w_i \, K(x, x_i), \tag{2.2}$$

where $\{x_i\}_{i=1}^n$ are the n observed input or explanatory variables. The statistical learning community as well as the approximation theory community has studied and characterized properties of the RKHS for various classes of kernels [DeVore *et al.* 1989, Zhou 2003].

Probabilistic versions and interpretations of kernel estimators have been of interest going back to the work of Hájek [1961, 1962] and Kimeldorf and Wahba [1971]. More recently a variety of kernel models with a Bayesian framework applied to the finite representation from the representer theorem have been proposed [Tipping 2001, Sollich 2002, Chakraborty *et al.* 2005]. However, the direct adoption of the finite representation is not a fully Bayesian model since it depends on the (arbitrary) training data sample size (see remark 2.1.3 for more discussion). In addition, this prior distribution is supported on a finite-dimensional subspace of the RKHS. Our coherent fully Bayesian approach requires the specification of a prior distribution over the entire space $\mathcal{H}$. A continuous, positive semi-definite kernel on a compact space $\mathcal{X}$ is called a *Mercer* kernel. The RKHS for such a kernel can be characterized [Mercer 1909, König 1986] as

$$\mathcal{H}_K = \left\{ f \;\middle|\; f(x) = \sum_{j \in \Lambda} a_j \phi_j(x) \text{ with } \sum_{j \in \Lambda} {a_j}^2 / \lambda_j < \infty \right\}, \tag{2.3}$$

where $\{\phi_j\} \subset \mathcal{H}$ and $\{\lambda_j\} \subset \mathbb{R}_+$ are the orthonormal eigenfunctions and the corresponding non-increasing eigenvalues of the integral operator with kernel K on

$L^2\big(\mathcal{X}, \mu(\mathrm{d}u)\big)$,

$$\lambda_j \phi_j(x) = \int_{\mathcal{X}} K(x,u)\phi_j(u)\mu(\mathrm{d}u) \tag{2.4}$$

and where $\Lambda := \{j : \lambda_j > 0\}$. The eigenfunctions $\{\phi_j\}$ and the eigenvalues $\{\lambda_j\}$ depend on the measure $\mu(\mathrm{d}u)$, but the RKHS does not. This suggests specifying a prior distribution over $\mathcal{H}$ by placing one on the parameter space

$$\mathcal{A} = \Big\{ \{a_j\} \,\Big|\, \sum_{j\in\Lambda} {a_j}^2/\lambda_j < \infty \Big\}$$

as in Johnstone [1998] and Wasserman [2005, Section 7.2]. There are serious computational and conceptual problems with specifying a prior distribution on this infinite-dimensional set. In particular, only in special cases are the eigenfunctions $\{\phi_j\}$ and eigenvalues $\{\lambda_j\}$ available in closed form.

Another approach, the *Bayesian kernel model*, is to study the class of functions expressible as kernel integrals

$$\mathcal{G} = \left\{ f \;\middle|\; f(x) = \int_{\mathcal{X}} K(x,u)\ \gamma(\mathrm{d}u), \quad \gamma \in \Gamma \right\}, \tag{2.5}$$

for some space $\Gamma \subseteq \mathscr{B}(\mathcal{X})$ of signed Borel measures. Any prior distribution on Γ induces one on $\mathcal{G}$. The natural question that arises in this Bayesian approach is:

> For what spaces Γ of signed measures is the RKHS $\mathcal{H}_K$ identical to the linear space $\mathrm{span}(\mathcal{G})$ spanned by the Bayesian kernel model?

The space $\mathcal{G}$ is the range $\mathscr{L}_K[\Gamma]$ of the integral operator $\mathscr{L}_K : \Gamma \to \mathcal{G}$ given by

$$\mathscr{L}_K[\gamma](x) := \int_{\mathcal{X}} K(x,u)\,\gamma(\mathrm{d}u). \tag{2.6}$$

Informally (we will be more precise in Section 2.2) we can characterize Γ as the range of the inverse operator $\mathscr{L}_K^{-1} : \mathcal{H}_K \to \Gamma$. The requirements on Γ for the equivalence

between $\mathscr{L}_K[\Gamma]$ and $\mathcal{H}_K$ is the primary focus of this paper and in Section 2.2 we formalize and prove the following proposition:

Proposition 2.1.1. *For $\Gamma = \mathscr{B}(\mathcal{X})$, the space of all signed Borel measures, $\mathcal{G} = \mathcal{H}_K$.*

The proposition asserts that the Bayesian kernel model and the penalized loss model both operate in the same function space when Γ includes all signed measures.

This result lays a theoretical foundation from a function analytic perspective for the use of two commonly used prior specifications: Dirichlet process priors [Ferguson 1973, West 1992, Escobar and West 1995, MacEachern and Müller 1998, Müller *et al.* 2004] and Lévy process priors [Wolpert *et al.* 2003, Wolpert and Ickstadt 2004].

2.1.1 Overview

In Section 2.2, we formalize and prove the above proposition. Prior distributions are placed on the space of signed measures in Section 2.4 using Lévy, Dirichlet, and Gaussian processes. In Section 2.5 we provide two examples using slightly different process prior distributions for a univariate regression problem and a high dimensional classification problem. This illustrates the use of these process priors for posterior inference. We close in Section 2.6 with a brief discussion.

Remark 2.1.2. *equation (2.5) is a Fredholm integral equation of the first kind [Fredholm 1900]. The problem of estimating an unknown measure γ for a specified element $f \in \mathcal{H}_K$ is ill-posed [Hadamard 1902] in the sense that small changes in f may give rise to large changes in estimates of γ. It was precisely the study of this problem that led Tikhonov [1963] to his regularization method, in a study of problems in numerical analysis such as interpolation or Gauss quadrature. Bayesian methods for interpolation and Gauss quadrature were explored by Diaconis [1988]. A Bayesian method using Lévy process priors to address numerically ill-posed problems was developed by*

Wolpert and Ickstadt [2004]. We will return to this relation between robust statistical estimation and numerically stable methods in the discussion.

Remark 2.1.3. *Due to the relation between regularization and Bayes estimators the finite representation is a MAP (maximal a posterior) estimator [Wahba 1999, Poggio and Girosi 1990]. However, functions in the RKHS having a posterior probability very close to that of the MAP estimator need not have a finite representation so building a prior only on the finite representation is problematic if one wants to estimate the full posterior on the entire RKHS. Thus the prior used to derive the MAP estimate is essentially the same as those used in Sollich [2002]. This will lead to serious computational and conceptual difficulties when the full posterior must be simulated.*

2.2 Characterizing the function space of the kernel model

In this section we formalize the relationship between the RKHS and the function space induced by the Bayesian kernel model.

2.2.1 Properties of the RKHS

Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous, positive semi-definite (*Mercer*) kernel. Consider the space of functions

$$\mathcal{H} = \left\{ f \;\middle|\; f(x) = \sum_{j=1}^{n} a_j K(x, x_j) : \; n \in \mathbb{N}, \; \{x_j\} \subset \mathcal{X}, \; \{a_j\} \subset \mathbb{R} \right\}$$

with an inner product $\langle \cdot \, , \, \cdot \rangle_K$ extending

$$\big\langle K(\cdot, x_i), \; K(\cdot, x_j) \big\rangle_K := K(x_i, x_j).$$

The Hilbert space closure $\mathcal{H}_K$ of $\mathcal{H}$ in this inner-product is the RKHS associated with the kernel K [Aronszajn 1950]. The kernel is "reproducing" in the sense that

each $f \in \mathcal{H}_K$ satisfies

$$f(x) = \langle f, K_x \rangle_K$$

for all $x \in \mathcal{X}$, where $K_x(\cdot) := K(\cdot, x)$.

A well-known alternate representation of the RKHS is given by an orthonormal expansion (Aronszajn 1950, extended to arbitrary measures by König 1986; see Cucker and Smale 2001). Let $\{\lambda_j\}$ and $\{\phi_j\}$ be the non increasing eigenvalues and corresponding complete orthonormal set of eigenvectors of the operator $\mathscr{L}_K$ of equation (2.6), restricted to the Hilbert space $L^2\big(\mathcal{X}, \mathrm{d}u\big)$ of measures $\gamma(\mathrm{d}u) = \gamma(u)\mathrm{d}u$ with square-integrable density functions $\gamma \in L^2\big(\mathcal{X}, \mathrm{d}u\big)$. Mercer's theorem [Mercer 1909] asserts the uniform and absolute convergence of the series

$$K(u, v) = \sum_{j=1}^{\infty} \lambda_j \, \phi_j(u)\phi_j(v), \tag{2.7}$$

whereupon with $\Lambda := \{j : \lambda_j > 0\}$ we have

$$\mathcal{H}_K = \left\{ f = \sum_{j \in \Lambda} a_j \phi_j \;\Big|\; \sum_{j \in \Lambda} {\lambda_j}^{-1} \, {a_j}^2 < \infty \right\}.$$

2.2.2 Bayesian kernel models and integral operators

Recall the Bayesian kernel model was defined by

$$\mathcal{G} = \left\{ \mathscr{L}_K[\gamma](x) := \int_{\mathcal{X}} K(x, u) \;\; \gamma(\mathrm{d}u), \quad \gamma \in \Gamma \right\},$$

where Γ is a space of signed Borel measures on $\mathcal{X}$. We wish to characterize the space $\mathscr{L}_K^{-1}(\mathcal{H}_K)$ of Borel measures mapped into the RKHS $\mathcal{H}_K$ of equation (2.3). A precise characterization is difficult and instead we will find a subclass $\Gamma \subset \mathscr{L}_K^{-1}(\mathcal{H}_K)$ which will be large enough in practice, in the sense that $\mathscr{L}_K(\Gamma)$ is dense in $\mathcal{H}_K$.

First we study the image under $\mathscr{L}_K$ of four classes of measures: (1) those with square integrable (Lebesgue) density functions; (2) all finite measures with Lebesgue density functions; (3) the set of discrete measures; and (4) linear combinations of all of these. Then we will extend these results to the general case of Borel measures.

We first examine the class $L^2(\mathcal{X}, \mathrm{d}u)$, viewed as the space of finite measures on $\mathcal{X}$ with square-integrable density functions with respect to Lebesgue measure; in a slight abuse of notation we write $\gamma(\mathrm{d}u) = \gamma(u)\mathrm{d}u$, using the same letter γ for the measure and its density function. Since $\mathcal{X}$ is compact and K bounded, $\mathscr{L}_K$ is a positive compact operator on $L^2(\mathcal{X}, \mathrm{d}u)$ with a complete ortho-normal system (CONS) $\{\phi_j\}$ of eigenfunctions with non increasing eigenvalues $\{\lambda_j\} \subset \mathbb{R}_+$ satisfying equation (2.7). Each $\gamma \in L^2(\mathcal{X}, \mathrm{d}u)$ admits a unique expansion $\gamma = \sum_j a_j\phi_j$, with $\|\gamma\|_2^2 = \sum_j a_j^2 < \infty$. The image under $\mathscr{L}_K$ of the measure $\gamma(\mathrm{d}u) := \gamma(u)\,\mathrm{d}u$ with Lebesgue density function γ may be expressed as the L^2-convergent sum

$$\mathscr{L}_K[\gamma](x) = \sum_j \lambda_j\, a_j\phi_j(x).$$

Proposition 2.2.1. *For every $\gamma \in L^2(\mathcal{X}, \mathrm{d}u)$, $\mathscr{L}_K[\gamma] \in \mathcal{H}_K$ and*

$$\|\mathscr{L}_K[\gamma]\|_K^2 = \langle \mathscr{L}_K[\gamma], \gamma\rangle_2.$$

Consequently, $L^2(\mathcal{X}, \mathrm{d}u) \subset \mathscr{L}_K^{-1}(\mathcal{H}_K)$.

Proof. It holds that

$$\|\mathscr{L}_K[\gamma]\|_K^2\| = \|\sum_{j\in\Lambda} \lambda_j a_j\phi_j\|_K^2 = \sum_{j\in\Lambda} \frac{(\lambda_j a_j)^2}{\lambda_j} = \sum_{j\in\Lambda} \lambda_j a_j^2$$

which is upper bounded by $\lambda_1 \sum_j a_j^2 < \infty$. Hence $\mathscr{L}_K[\gamma] \in \mathcal{H}_K$. By direct computation, we have

$$\langle \mathscr{L}_K[\gamma], \gamma\rangle_2 = \Big\langle \sum \lambda_j a_k\phi_j, \sum a_j\phi_j \Big\rangle_2 = \sum \lambda_k a_j^2 = \|\mathscr{L}_K[\gamma]\|_K^2.$$

■

The following corollary illustrates that the space $L^2(\mathcal{X}, \mathrm{d}u)$ is too small for our purpose–*i.e.*, that important functions $f \in \mathscr{L}_K^{-1}(\mathcal{H}_K)$ fail to lie in $L^2(\mathcal{X}, \mathrm{d}u)$.

Corollary 2.2.2. *If the set $\Lambda := \{j : \lambda_j > 0\}$ is finite, then $\mathscr{L}_K(L^2(\mathcal{X}, \mathrm{d}u)) = \mathcal{H}_K$; otherwise $\mathscr{L}_K(L^2(\mathcal{X}, \mathrm{d}u)) \subsetneqq \mathcal{H}_K$. The latter occurs whenever K is strictly positive definite and the RKHS is infinite-dimensional.*

Proof. The first claim is obvious since both $\mathscr{L}_K[L^2(\mathcal{X}, \mathrm{d}u)]$ and $\mathcal{H}_K$ are the same finite dimensional space spanned by $\{\phi_j\}_{j\in\Lambda}$.

The second claim follows from the existence of the sequence $(c_j)_{j\in\Lambda}$ such that

$$\sum_{j\in\Lambda} \frac{c_j^2}{\lambda_j} < \infty \qquad \text{and} \qquad \sum_{j\in\Lambda} \frac{c_j^2}{\lambda_j^2} = \infty.$$

For any such sequence, the function $f = \sum_{j\in\Lambda} c_j\phi_j$ lies in $\mathcal{H}_K$. But by Proposition 2.2.1, one cannot find a $\gamma \in L^2(\mathcal{X}, \mathrm{d}u)$ such that $\mathscr{L}_K[\gamma] = f$. A simple example is $(c_j)_{j\in\Lambda} = (\lambda_j)_{j\in\Lambda}$.

If K is strictly positive definite, then all its eigenvalues are positive. So the last claim holds. ■

Thus only for finite dimensional RKHS's is the space of square integrable functions sufficient to span the RKHS. In almost all interesting non-parametric statistics problems, the RKHS is infinite-dimensional.

Next we examine the space of integrable functions $L^1(\mathcal{X}, \mathrm{d}u)$, a larger space than $L^2(\mathcal{X}, \mathrm{d}u)$ when $\mathcal{X}$ is compact.

Proposition 2.2.3. *For every $\gamma \in L^1(\mathcal{X}, \mathrm{d}u)$, $\mathscr{L}_K[\gamma] \in \mathcal{H}_K$. Consequently, $L^1(\mathcal{X}, \mathrm{d}u) \subset \mathscr{L}_K^{-1}(\mathcal{H}_K)$.*

Proof. Since $K(u,v)$ is continuous on the compact set $\mathcal{X}\times\mathcal{X}$, it has a finite maximum $\kappa^2 := \sup_{u,v} K(u,v) < \infty$. Since $L^2(\mathcal{X},\mathrm{d}u)$ is dense in $L^1(\mathcal{X},\mathrm{d}u)$, for every $\gamma\in L^1(\mathcal{X},\mathrm{d}u)$, there exists a Cauchy sequence $\{\gamma_n\}_{n\geq 1}\subset L^2(\mathcal{X},\mathrm{d}u)$ which converges to γ in $L^1(\mathcal{X},\mathrm{d}u)$. It follows from Proposition 2.2.1 that $\mathscr{L}_K[\gamma_n]\in\mathcal{H}_K$ and

$$\|\mathscr{L}_K[\gamma_n]\|_K^2 = \int_{\mathcal{X}}\int_{\mathcal{X}} K(u,v)\,\gamma_n(u)\,\mathrm{d}u\,\gamma_n(v)\,\mathrm{d}v \leq \kappa^2\int_{\mathcal{X}}|\gamma_n(u)|\mathrm{d}u\int_{\mathcal{X}}|\gamma_n(v)|\mathrm{d}v = \kappa^2\|\gamma_n\|_1^2 < \infty.$$

Therefore we have $\{\mathscr{L}_K[\gamma_n]\}_{n\geq 1}\subset\mathcal{H}_K$ and

$$\lim_{n\to\infty}\sup_{m>n}\|\mathscr{L}_K[\gamma_n]-\mathscr{L}_K[\gamma_m]\|_K \leq \lim_{n\to\infty}\sup_{m>n}\kappa\|\gamma_n-\gamma_m\|_1 = 0,$$

so $\{\mathscr{L}_K[\gamma_n]\}_{n\geq 1}$ is a Cauchy sequence in $\mathcal{H}_K$. By completeness it converges to some $f\in\mathcal{H}_K$. The proof will be finished if we show $\mathscr{L}_K[\gamma]=f$.

By the reproducing property of $\mathcal{H}_K$ convergence in the RKHS norm implies pointwise convergence for $x\in\mathcal{X}$, so $L_K[\gamma_n](x)\to f(x)$ for every x.

In addition, for every $x\in\mathcal{X}$, we have

$$\lim_{n\to\infty}|\mathscr{L}_K[\gamma_n](x)-\mathscr{L}_K[\gamma](x)| \leq \int_{\mathcal{X}}|K(x,u)(\gamma_n(u)-\gamma(u))|\,\mathrm{d}u \leq \kappa^2\|\gamma_n-\gamma\|_1 = 0,$$

which implies that $\mathscr{L}_K[\gamma_n](x)$ also converges to $\mathscr{L}_K[\gamma](x)$. Hence $\mathscr{L}_K[\gamma]=f\in\mathcal{H}_K$. ■

Another class of functions to be considered is the space of finite discrete measures,

$$\mathscr{M}_D = \left\{\mu=\sum_j c_j\delta_{x_j}:\ \{c_j\}\subset\mathbb{R},\ \{x_j\}\subset\mathcal{X},\ \sum_j|c_j|<\infty\right\},$$

where δ_x is the Dirac measure supported at $x\in\mathcal{X}$ (the sum may be finite or infinite). This class will arise naturally when we examine Lévy and Dirichlet processes in Section 2.4.3.

Proposition 2.2.4. *For every $\mu \in \mathscr{M}_D$, $\mathscr{L}_K[\mu] \in \mathcal{H}_K$. Consequently, $\mathscr{M}_D \subset \mathscr{L}_K^{-1}(\mathcal{H}_K)$.*

Proof. Let $\gamma = \sum c_i \delta_{x_i} \in \mathscr{M}_D$. Then $\mathscr{L}_K[\gamma] = \sum c_i K_{x_i}$ and

$$\|\mathscr{L}_K[\gamma]\|_K^2 = \sum_{i,j} c_i K(x_i, x_j) c_j \leq \kappa^2 \left(\sum_i |c_i| \right)^2 < \infty.$$

Therefore, our conclusion holds. ■

By Proposition 2.2.3 and 2.2.4 the space $\mathscr{M}$ spanned by $L^1(\mathcal{X}, \mathrm{d}u) \cup \mathscr{M}_D$ is a subset of $\mathscr{L}_K^{-1}(\mathcal{H}_K)$. The range of $\mathscr{L}_K$ on just the elements of $\mathscr{M}_D$ with *finite* support is precisely $\mathcal{H}$, linear combinations of the $\left\{K_{x_j}\right\}_{x_j \in \mathcal{X}}$; thus

Proposition 2.2.5. *$\mathscr{L}_K(\mathscr{M})$ is dense in $\mathcal{H}_K$ with respect to the RKHS norm.*

Let $\mathscr{B}_+(\mathcal{X})$ denote the cone of all finite nonnegative Borel measures on $\mathcal{X}$ and $\mathscr{B}(\mathcal{X})$ the set of signed Borel measures. Clearly every $\mu \in \mathscr{B}(\mathcal{X})$ can be written uniquely as $\mu = \mu_+ - \mu_-$ with $\mu_+, \mu_- \in \mathscr{B}_+(\mathcal{X})$. The set $\mathscr{B} \backslash \mathscr{M}$ contains those Borel measures that are singular with respect to the Lebesgue measure. In this context, the set $\mathscr{M} = \mathscr{M}_D \cup L^1(\mathcal{X}, \mathrm{d}u)$ contains the Borel measures that can be used in practice. The above results, Propositions 2.2.3 and 2.2.1, also hold if we replace the Lebesgue measure with a Borel measure. It is natural to compare $\mathscr{B}(\mathcal{X})$ with $\mathscr{L}_K^{-1}(\mathcal{H}_K)$.

Proposition 2.2.6. *$\mathscr{B}(\mathcal{X}) \subset \mathscr{L}_K^{-1}(\mathcal{H}_K)$.*

Proof. The arguments for Lebesgue measure hold if we replace the Lebesgue measure with any finite Borel measure. We denote the corresponding integral operator as $\mathcal{L}_{K,\mu}$ and function space of integrable and square integrable functions as $L^1_\mu(\mathcal{X})$ and $L^2_\mu(\mathcal{X})$ respectively. Then

$$L^2_\mu(\mathcal{X}) \subset L^1_\mu(\mathcal{X}) \subset L^{-1}_{K,\mu}(\mathcal{H}_K).$$

Since the function $1_{\mathcal{X}}(x) = 1$ lies in $L^1_\mu(\mathcal{X})$ we obtain

$$\mathscr{L}_{\scriptscriptstyle K}(\mu) = \mathcal{L}_{K,\mu}(1_{\mathcal{X}}) = \int_{\mathcal{X}} K(\cdot, u)\mathrm{d}\mu(u) \in \mathcal{H}_K.$$

This implies $\mathscr{B}_+(\mathcal{X})$ lies in $L_K^{-1}(\mathcal{H}_K)$ and so does $\mathscr{B}(\mathcal{X})$. ■

We close this section by showing that even $\mathscr{B}(\mathcal{X})$ need not exactly characterize the class $\mathscr{L}_{\scriptscriptstyle K}^{-1}(\mathcal{H}_K)$. The proof of Proposition 2.2.3 implies that

$$\|\mathscr{L}_{\scriptscriptstyle K}[\gamma]\|_K^2 = \iint_{\mathcal{X}\times\mathcal{X}} K(x,u)\,\gamma(x)\,\gamma(u)\,\mathrm{d}x\,\mathrm{d}u. \tag{2.8}$$

From the above it is apparent that $\mathscr{L}_{\scriptscriptstyle K}[\gamma] \in \mathcal{H}_K$ holds only if $\mathscr{L}_{\scriptscriptstyle K}[\gamma]$ is well defined and the quantity on the right hand side of (2.8) is finite. If the kernel is smooth and vanishes at certain $x, u \in \mathcal{X}$, then (2.8) can be finite even if $\gamma \notin L^1(\mathcal{X}, \mathrm{d}u)$. For example in the case of polynomial kernels δ'_x, the functional derivatives of the Dirac measure δ_x, are mapped into $\mathcal{H}_K$.

Proposition 2.2.7. $\mathscr{B}(\mathcal{X}) \subsetneqq \mathscr{L}_{\scriptscriptstyle K}^{-1}(\mathcal{H}_K(\mathcal{X}))$.

Proof. We construct an infinite signed measure γ satisfying $\mathscr{L}_{\scriptscriptstyle K}[\gamma] \in \mathcal{H}_K$. As in Example 1 below, let

$$K(x,u) := x \wedge u - xu$$

be the covariance kernel for the Brownian bridge on the unit interval $\mathcal{X} = [0,1]$ (as usual, "$x \wedge u$" denotes the minimum of two real numbers x, u). Consider the improper $\mathbf{B}e(0,0)$ distribution

$$\gamma(\mathrm{d}u) = \frac{\mathrm{d}u}{u(1-u)},$$

with image under the integral operator

$$f(x) := \mathscr{L}_{\scriptscriptstyle K}[\gamma](x) = -x\log(x) - (1-x)\log(1-x).$$

The function $f(x)$ satisfies $f(0) = 0 = f(1)$ and has finite RKHS norm

$$\|f\|_K^2 = -2\int_0^1 \frac{\log(x)}{1-x}\,\mathrm{d}x = \frac{\pi^2}{3},$$

so $f(x)$ is in the the RKHS (see Example 1). Thus the infinite signed measure $\gamma(\mathrm{d}s)$ is in $\mathscr{L}_K^{-1}[\mathcal{H}_K]$ but not in $\mathscr{B}(\mathcal{X})$, so $\mathscr{L}_K^{-1}[\mathcal{H}_K]$ is larger than the space of finite signed measures. ■

2.3 Two Concrete Examples

In this section we construct two explicit examples to help illustrate the ideas of Section 2.2.

Example 2.3.1 (Brownian bridge)**.** *On the space $\mathcal{X} = [0,1]$ consider the kernel*

$$K(x,u) := (x \wedge u) - xu,$$

which is jointly continuous and the covariance function for the Brownian bridge [Rogers and Williams 1987, §IV.40] and hence a Mercer kernel. The eigenfunctions and eigenvalues of equation (2.4) for Lebesgue measure $\mu(\mathrm{d}u) = \mathrm{d}u$ are

$$\lambda_j = \frac{1}{j^2\pi^2} \qquad \phi_j(x) = \sqrt{2}\,\sin(j\pi x).$$

The associate integral operator of equation (2.6) is

$$\begin{aligned}\mathscr{L}_K[\gamma](x) &:= \int_{\mathcal{X}} K(x,u)\,\gamma(\mathrm{d}u)\\ &= (1-x)\int_{[0,x)} u\gamma(\mathrm{d}u) + x\int_{[x,1]}(1-u)\gamma(\mathrm{d}u),\end{aligned}$$

mapping any $\gamma(\mathrm{d}u) = \gamma(u)\mathrm{d}u$ with $\gamma \in L^1(\mathcal{X}, \mathrm{d}u)$ to a function $f(x) = \mathscr{L}_K[\gamma](x)$ that satisfies the boundary conditions $f(0) = 0 = f(1)$ and, for almost every $x \in \mathcal{X}$,

$$
\begin{aligned}
f(x) &= (1-x)\int_0^x u\gamma(u)\mathrm{d}u + x\int_x^1 (1-u)\gamma(u)\mathrm{d}u \\
f'(x) &= \int_x^1 \gamma(u)\,\mathrm{d}u - \int_0^1 u\,\gamma(u)\,\mathrm{d}u \\
f''(x) &= -\gamma(x)
\end{aligned}
$$

and hence, by equation (2.8) and integration by parts,

$$
\begin{aligned}
\|f\|_K^2 &= \int_0^1 f(x)\,\gamma(x)\,\mathrm{d}x \\
&= \int_0^1 -f(x)\,f''(x)\,\mathrm{d}x \\
&= \int_0^1 f'(x)^2\,\mathrm{d}x.
\end{aligned}
$$

Evidently the RKHS is just

$$
\begin{aligned}
\mathcal{H}_K &= \left\{ f(x) = \sum_{j=1}^{\infty} a_j \sqrt{2}\sin(j\pi x) \;\Big|\; \sum_{j=1}^{\infty} \pi^2\, j^2\, {a_j}^2 < \infty \right\} \\
&= \left\{ f \text{ in } L^2(\mathcal{X},\mathrm{d}u) \big|\; f(0) = 0 = f(1) \text{ and } f' \in L^2(\mathcal{X},\mathrm{d}u) \right\},
\end{aligned}
$$

the subspace of the Sobolev space $H_{+1}(\mathcal{X})$ satisfying Dirichlet boundary conditions [Mazja 1985, Section 1.1.4], and

$$
\begin{aligned}
\mathscr{L}_K^{-1}\big(\mathcal{H}_K\big) &= \left\{ \gamma(x) = \sum_{j=1}^{\infty} a_j \sqrt{2}\sin(j\pi x) \;\Big|\; \sum_{j=1}^{\infty} \frac{{a_j}^2}{\pi^2 j^2} < \infty \right\} \\
&= \left\{ \gamma = f'' \big| f, f' \in L^2(\mathcal{X},\mathrm{d}u),\; f(0) = 0 = f(1) \right\},
\end{aligned}
$$

a subspace of $H_{-1}(\mathcal{X})$.

Example 2.3.2 (Splines on a circle)**.** *The kernel function for first order splines on the real line is*

$$
K(x,u) := |x-u| \qquad x, u \in \mathbb{R}
$$

and the corresponding RKHS norm is

$$\|f\|_K^2 = \int_{-\infty}^{\infty} f'(x)^2 \, \mathrm{d}x.$$

However, since the domain is not compact the spectrum of the associated integral operator on $L^2(\mathbb{R}, \mathrm{d}u)$ *is continuous rather than discrete, the approach of Section 2.2 does not apply.*

Instead we consider the case of splines with periodic boundary conditions. On the space $\mathcal{X} = [0, 1]$ *we consider the kernel function*

$$\begin{aligned} K(x,u) &= \sum_{j=1}^{\infty} \frac{1}{2\pi^2 j^2} \cos(2\pi j|u-x|) \\ &= \frac{1}{2}\Big(|x-u| - \frac{1}{2}\Big)^2 - \frac{1}{24} \qquad 0 < x, u < 1 \end{aligned}$$

The eigenfunctions and eigenvalues of equation (2.4) for Lebesgue measure $\mu(\mathrm{d}u) = \mathrm{d}u$ *are*

$$\begin{array}{lllll} \phi_{2j-1}(x) := \sqrt{2}\sin(2\pi j x) & & \lambda_{2j-1} = \frac{1}{4\pi^2 j^2} & & \\ \phi_{2j}(x) := \sqrt{2}\cos(2\pi j x) & & \lambda_{2j} = \frac{1}{4\pi^2 j^2} & & j \in \mathbb{N}. \end{array}$$

The RKHS norm for this kernel is

$$\|f\|_K^2 = \int_0^1 f'(x)^2 \, \mathrm{d}x$$

and the RKHS is

$$\mathcal{H}_K = \left\{ f(x) = \sum_{j=1}^{\infty} \sqrt{2}\big[a_j \sin(2\pi j x) + b_j \cos(2\pi j x)\big] \;\Big|\; \sum_{j=1}^{\infty} 4\pi^2 j^2 (a_j^2 + b_j^2) < \infty \right\}$$

the subspace of the Sobolev space $H_{+1}(\mathcal{X})$ *satisfying periodic boundary conditions and orthogonal to the constants [Wahba 1990, Section 2.1] and*

$$\mathscr{L}_K^{-1}(\mathcal{H}_K) = \left\{ \gamma(x) = \sum_{j=1}^{\infty} \sqrt{2}\big[a_j \sin(\pi j x) + b_j \cos(\pi j x)\big] \;\Big|\; \sum_{j=1}^{\infty} \frac{a_j^2 + b_j^2}{4j^2\pi^2} < \infty \right\},$$

a subspace of $H_{-1}(\mathcal{X})$.

Elements in either RKHS given in the above two examples with a finite representation

$$f(x) = \sum_{i=1}^{m} c_i K(x, x_i), \;\; m < \infty$$

are splines. For the first example these functions are linear splines that vanish at $\{0, 1\}$. In the second example if the coefficients sum to zero ($\sum_{i=1}^{m} c_i = 0$), then these functions are linear splines with periodic boundary conditions. If the coefficients do not sum to zero then they are quadratic splines with periodic boundary conditions.

2.4 Bayesian Kernel Models

Our goal from Section 2.1 is to present a coherent Bayesian framework for non-parametric function estimation in a RKHS. Suppose we observe data (with noise), $\{(x_i, y_i)\} \subset \mathcal{X} \times \mathbb{R}$ from the linear regression model

$$y_i = f(x_i) + \varepsilon_i \tag{2.9}$$

where we assume $\{\varepsilon_i\}$ are independent $\mathbf{No}(0, \sigma^2)$ random variables with unknown variance σ^2, and $f(\cdot)$ is an unknown function we wish to estimate. For a fixed kernel we assume $f \in \mathcal{H}_K$. Recall that the integral operator $\mathscr{L}_K$ maps $\mathscr{M}(\mathcal{X})$ into $\mathcal{H}_K$ and in particular $\mathscr{L}_K(\mathscr{M}(\mathcal{X}))$ is dense in $\mathcal{H}_K$. Therefore, we assume that

$$f(x) = \int_{\mathcal{X}} K(x, u) Z(du) \tag{2.10}$$

where $Z(du) \in \mathscr{M}(\mathcal{X})$ is a signed measure on $\mathcal{X}$. If we put a prior on $\mathscr{M}(\mathcal{X})$, we are in essence putting a prior on the functions $f \in \mathcal{G}$.

Our measurement error model (2.9) gives us the following likelihood for the data $D := \{(x_i, y_i)\}_{i=1}^n$

$$L(D|Z) \propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - f(x_i)\right)^2\right]. \tag{2.11}$$

With a prior distribution on Z, $\pi(Z)$, we can obtain the posterior density function given data

$$\pi(Z|D) \propto L(D|Z)\ \pi(Z), \tag{2.12}$$

which implies a posterior distribution for f via the integral operator (2.10).

2.4.1 Priors on $\mathscr{M}$

A random signed measure $Z(du)$ on $\mathcal{X}$ can be viewed as a stochastic process on $\mathcal{X}$. Therefore the practice of specifying a prior on $\mathscr{M}(\mathcal{X})$ via a stochastic process is ubiquitous in non-parametric Bayesian analysis. Gaussian processes and Dirichlet processes are two commonly used stochastic processes to generate random measures.

We first apply the results of Section 2.2 to Gaussian process priors [Rasmussen and Williams 2006, Section 6] and then to Lévy process priors [Wolpert *et al.* 2003, Tu *et al.* 2006]. We also remark that Dirichlet processes can be constructed from Lévy process priors.

2.4.2 Gaussian Processes

Gaussian processes are canonical examples of stochastic processes used for generating random measures. They have been used extensively in the machine learning and statistics community with promising results in practice and theory [Kimeldorf and Wahba 1971, Chakraborty *et al.* 2005, Rasmussen and Williams 2006, Ghosal and Roy 2006].

We consider two modeling approaches using Gaussian process priors:

i. Model I: Placing a prior directly on the space of functions $f(x)$ by sampling from paths of the Gaussian process with its covariance structure defined via a kernel K;

ii. Model II: Placing a prior on the random signed measures $Z(du)$ on $\mathcal{X}$ by using a Gaussian process prior for $Z(du)$ which implies a prior on the function space defined by the kernel model in equation (2.10).

For both approaches we can characterize the function space spanned by the kernel model. The first approach is the more standard approach for non-parametric Bayesian inference using Gaussian processes while the later is an example of our Bayesian kernel model. However, as pointed out by [Wahba 1990, Section 1.4] the random functions from the first approach will be almost surely outside the RKHS induced by the kernel. However these functions will be contained in a larger RKHS, as we show in the next section.

We first state some classical results on the sample paths of Gaussian processes. We then use these properties and the results of Section 2.2 to characterize the function spaces of the two models.

Sample paths of Gaussian processes

Consider a Gaussian process $\{Z_u,\ u \in \mathcal{X}\}$ on a probability space $\{\Omega, \mathcal{A}, \mathbb{P}\}$ having covariance functions determined by a kernel function K. Let $\mathcal{H}_K$ be the corresponding RKHS and let the mean m be contained in the RKHS, $m \in \mathcal{H}_K$. Then the following zero-one law holds:

Theorem 2.4.1. *(Kallianpur [1970], Theorem 5.1) If $Z_\bullet \equiv \{Z_u, u \in \mathcal{X}\}$ is a Gaussian process with covariance K and mean $m \in \mathcal{H}_K$, and $\mathcal{H}_K$ is infinite dimensional, then*

$$\mathbb{P}(Z_\bullet \in \mathcal{H}_K) = 0.$$

The probability measure is assumed to be complete.

Thus the sample paths of the Gaussian process are almost surely outside $\mathcal{H}_K$. However, there exists a RKHS $\mathcal{H}_R$ that is bigger than $\mathcal{H}_K$ that contains the sample paths almost surely. To construct such an RKHS we first need to define nuclear dominance.

Definition 2.4.2. *Given two kernel functions R and K, R dominates K (written as $R \succ K$) if $\mathcal{H}_K \subseteq \mathcal{H}_R$.*

Given the above definition of dominance the following operator can be defined:

Theorem 2.4.3. *[Lukić and Beder 2001] Let $R \succ K$. Then*

$$\|g\|_R \leq \|g\|_K, \ \ \forall g \in \mathcal{H}_K.$$

There exists a unique linear operator $L : \mathcal{H}_R \to \mathcal{H}_R$ whose range is contained in $\mathcal{H}_K$ such that

$$\langle f, g \rangle_R = \langle Lf, g \rangle_K, \ \ \forall f \in \mathcal{H}_R, \forall g \in \mathcal{H}_K.$$

In particular

$$LR_u = K_u, \ \ \ \forall u \in \mathcal{X}.$$

As an operator into $\mathcal{H}_R$, L is bounded, symmetric, and positive.

Conversely, let $L : \mathcal{H}_R \to \mathcal{H}_R$ be a positive, continuous, self-adjoint operator then

$$K(s,t) = \langle LR_s, R_t \rangle_R, \ \ s, t \in \mathcal{X}$$

defines a reproducing kernel on $\mathcal{X}$ such that $K \leq R$.

L is the dominance operator of $\mathcal{H}_R$ over $\mathcal{H}_K$ and this dominance is called nuclear if L is a nuclear or trace class operator (a compact operator for which a trace may be defined that is finite and independent of the choice of basis). We denote nuclear dominance as $R \twoheadrightarrow K$.

Implications for the function spaces of the models

Model I placed a prior directly on the space of functions using sample paths from the Gaussian process with covariance structure defined by the kernel K. Theorem 2.4.1 states that sample paths from this Gaussian process are not contained in $\mathcal{H}_K$. However, there exists another RKHS $\mathcal{H}_R$ with kernel R which does contain the sample path if R has nuclear dominance over K.

Theorem 2.4.4. *[Lukić and Beder 2001] Let K and R be two reproducing kernels. Assume that the RKHS $\mathcal{H}_R$ is separable. A necessary and sufficient condition for the existence of a Gaussian process with covariance K and mean $m \in \mathcal{H}_R$ and with trajectories in $\mathcal{H}_R$ with probability 1 is that $R \gtrdot K$.*

The implication of this theorem is that we can find a function space $\mathcal{H}_R$ that contains functions generated by the Gaussian process defined by covariance function K.

Model II places a prior on random signed measures $Z(du)$ on $\mathcal{X}$ by using a Gaussian process prior for $Z(du)$. This implies a prior of the space of functions spanned by the kernel model in equation (2.10). This space $\mathcal{G}$ is contained in $\mathcal{H}_K$ by our results in Section 2.2. This is due to the fact that any sample path from a continuous Gaussian process on a compact domain $\mathcal{X}$ is in L^1 and therefore the corresponding function from the integral (2.10) is still in $\mathcal{H}_K$.

2.4.3 Lévy processes

Lévy processes offer an alternative to Gaussian processes in non-parametric Bayesian modeling. Dirichlet processes and Gaussian processes with a particular covariance structure can be formulated from the framework of Lévy processes. For the sake of simplicity in exposition, we will use the univariate setting $\mathcal{X} = [0, 1]$ to illustrate the

construction of random signed measures using Lévy processes. The extension to the multivariate setting is straightforward and outlined in Appendix **??**.

A stochastic process $Z := \{Z_u \in \mathbb{R} : u \in \mathcal{X}\}$ is called a *Lévy process* if it satisfies the following conditions:

1. $Z_0 = 0$ almost surely.
2. For any integer $m \in \mathbb{N}$ and any $0 = u_0 < u_1 < ... < u_m$, the random variables $\{Z_{u_j} - Z_{u_{j-1}}\},\ 1 \leq j \leq m$ are independent. (Independent increments property)
3. The distribution of $Z_{s+u} - Z_s$ does not depend on s (Temporal homogeneity or stationary increments property).
4. The sample paths of Z are almost surely right continuous and have left limits, *i.e.*, are "cádlág".

Familiar examples of Lévy processes include Brownian motion, Poisson processes, and gamma processes. The following celebrated theorem characterizes Lévy processes.

Theorem 2.4.5. ***(Lévy-Khintchine)*** *Z is a Lévy process if and only if the characteristic function of $Z_u : u \geq 0$ has the following form:*

$$\mathbb{E}[e^{i\lambda Z_u}] = \exp\left\{u\left[i\lambda a - \frac{1}{2}\sigma^2\lambda^2 + \int_{\mathbb{R}\setminus 0}[e^{i\lambda w} - 1 - i\lambda w 1_{\{w:|w|<1\}}(w)]\nu(dw)\right]\right\}, \tag{2.13}$$

where $a \in \mathbb{R}$, $\sigma^2 \geq 0$ and ν is a nonnegative measure on $\mathbb{R}\setminus 0$ with

$$\int_{\mathbb{R}\setminus 0}(1 \wedge |w|^2)\nu(dw) < \infty. \tag{2.14}$$

Note that (2.13) can be written as a product of two components,

$$\exp\left\{iau\,\lambda - \frac{u\sigma^2}{2}\lambda^2\right\} \ \times \ \exp\left\{u\int_{\mathbb{R}\setminus 0}\left[e^{i\lambda w} - 1 - i\lambda w 1_{\{w:|w|<1\}}(w)\right]\nu(dw)\right\},$$

the characteristic functions of a Gaussian process and of a partially compensated Poisson process, respectively. This observation is the essence of the Lévy-Itô theorem [Applebaum 2004, Theorem 2.4.16], which asserts that every Lévy process can be decomposed into the sum of two independent components: a "continuous process" (Brownian motion with drift) and a (possibly compensated) "pure jump" process. The three parameters (a, σ^2, ν) in (2.13) uniquely determine a Lévy process where a denotes the drift term, σ^2 denotes the variance (diffusion coefficient) of the Brownian motion, and $\nu(dw)$ denotes the intensity of the jump process. The so-called "Lévy measure" ν need not be finite, but (2.14) implies that $\nu[(-\epsilon, \epsilon)^c] < \infty$ for each $\epsilon > 0$ and so ν is at least sigma-finite.

Pure jump Lévy processes

Pure jump Lévy processes are used extensively in non-parametric Bayesian statistics due to their computationally amenability. In this section we first state an interpretation of these processes using Poisson random fields. We then describe Dirichlet and symmetric α-stable processes.

Poisson random fields interpretation

Any pure jump Lévy process Z has a nice representation via a Poisson random field. Set $\Delta Z_u := Z_u - \lim_{s \uparrow u} Z_s$, the jump size at the location u. Set $\Gamma = \mathbb{R} \times \mathcal{X}$, the Cartesian product of $\mathbb{R}$ with $\mathcal{X}$. For any sets $A \subset \mathbb{R} \backslash 0$ bounded away from zero and $B \subset \mathcal{X}$ we can define the counting measure

$$N(A \times B) := \sum_{s \in B} 1_A\big(\Delta Z_s\big). \tag{2.15}$$

The measure N defined above turns out to be a Poisson random measure on Γ, with mean measure $\nu(dw)du$ where du is the uniform reference measure on $\mathcal{X}$ (for instance

the Lebesgue measure when $\mathcal{X} = [0, 1]$). For any $E \subset \Gamma$ with $\mu = \int_E \nu(dw)du < \infty$ the random variable $N(E)$ has a Poisson distribution with intensity μ.

When ν is a finite measure, the total number of jumps $J \in \mathbb{N}$ of the process follows a Poisson distribution with finite intensity $\mu(\Gamma)$. When Z has a density with respect to the Lévy random field M with Lévy measure m, Z_u has finite total variation and determines a finite measure $Z(du) = dZ_u$. In this case, any realization of $Z(du)$ can be formulated as

$$Z(du) = \sum_{j=1}^{J} w_j \delta_{u_j}, \tag{2.16}$$

where $(w_j, u_j) \in \Gamma$ are *i.i.d.* draws from $\nu(dw)du$ representing the jump size and the jump location, respectively. Given a realization of $Z(du) = \{u_j, w_j\}_{j=1}^{J}$, equation (2.10) reduces to

$$\int_{\mathcal{X}} K(x, u)Z(du) = \int_{\Gamma} K(x, u)N(dwdu) = \sum_{j=1}^{J} w_j K(x, u_j),$$

where $N(dwdu)$ is a Poisson random measure as defined by (2.15). Then the likelihood for the data $D := \{(x_i, y_i)\}_{i=1}^{n}$ is given by

$$L(D|Z) \propto \prod_{i=1}^{n} \exp\Big[-\frac{1}{2\sigma^2}\Big(y_i - \sum_{j=1}^{J} w_j K(x_i, u_j)\Big)^2\Big].$$

If the measure $\nu(dw)du$ has a density function $\nu(w, u)$ with respect to some finite reference measure $m(dwdu)$, then the prior density function for Z with respect to a Lévy(m) process is

$$\pi(Z) = \Big[\prod_{j=1}^{J} \nu(w_j, u_j)\Big] e^{m(\Gamma)-\nu(\Gamma)}. \tag{2.17}$$

Using Bayes' theorem, we can calculate the posterior distribution for Z via (2.12).

When ν is an infinite measure the number of jumps in the unit interval is countably infinite almost surely. However, if the Lévy measure satisfies

$$\int_{\mathbb{R}} (1 \wedge |w|)\nu(dw) < \infty, \tag{2.18}$$

then the sequence $\{w_j\}$ is almost surely absolutely summable (i.e, $\sum_{j=1}^{\infty} |w_j| < \infty$ a.s.) and we can still represent the process Z via the summation (2.16). Note that condition (2.18) is stronger than the integrability condition (2.14) in the Lévy-Khintchine theorem. This allows for the existence of Lévy processes with jumps that are not absolutely summable.

Dirichlet process

The Dirichlet process is commonly used in non-parametric Bayesian analysis [Ferguson 1973, 1974] mainly due to its analytical tractability. When passing from prior to posterior computations, it has been shown that the Dirichlet process is the only conjugate member of the whole class of normalized random measures with independent increments [James *et al.* 2005] so the posterior can be efficiently computed. Recently it has received much attention in the machine learning literature [Blei and Jordan 2006, Xing *et al.* 2004, 2006]. Though Dirichlet processes are often defined via Dirichlet distributions, they can also be defined as a normalized Gamma process as noted by Ferguson [1973]. A Gamma process is a pure jump Lévy process, which has the Lévy measure

$$\nu(dw) = aw^{-1} \exp\{-bw\}dw, \quad w > 0,$$

so at each location u $Z_u \sim \text{Gamma}(au, b)$. Suppose Z_u is a Gamma$(a, 1)$ process defined on $\mathcal{X} = [0, 1]$, then

$$\tilde{Z}_u = Z_u / Z_1$$

is the $\mathbf{DP}(a\,\mathrm{d}u)$ Dirichlet process. Since the Dirichlet process is a random measure on probability distribution functions, it can be used when the target function $f(x)$ is a probability density function. Dirichlet processes can also be used to model a general smooth function $f(x)$ in combination with other random processes. For example, Liang *et al.* [2007a] and Liang *et al.* [2007b] consider a variation of the integral (2.10)

$$f(x) = \int_{\mathcal{X}} K(x,u)\, Z(du) = \int_{\mathcal{X}} w(u)\, K(x,u)\, F(du), \tag{2.19}$$

where the random signed measure $Z(du)$ is modeled by a random probability distribution function $F(du)$ and random coefficients $w(u)$. A Dirichlet process prior is specified for F and a Gaussian prior distribution is specified for w.

Symmetric α-stable process

Symmetric α-stable processes are another class of Lévy processes, arising from symmetric α-stable distributions. The symmetric α-stable distribution has the following characteristic function:

$$\varphi(\eta) = \exp(-\gamma|\eta|^{\alpha}),$$

γ is the dispersion parameter, and $\alpha \in (0,2]$ is the characteristic exponent. The case, when $\gamma = 1$ is called the standard symmetric α-stable (SαS) distribution. It has the following Lévy measure

$$\nu(dw) = \frac{\Gamma(\alpha+1)}{\pi} \sin\left(\tfrac{\pi\alpha}{2}\right) |w|^{-1-\alpha} dw \qquad \alpha \in (0,2].$$

Two important cases of SαS distributions are the Gaussian when $\alpha = 2$ and the Cauchy when $\alpha = 1$. Thus SαS processes allow us to model heavy or light tail processes by varying α. One can verify that the Lévy measure is infinite for $0 < \alpha \leq 2$ since $\nu(\mathbb{R}) = \int_{\mathbb{R}} \nu(dw) = 2\int_{(0,\infty]} \alpha w^{-1-\alpha}\, dw = \infty$. Hence the process has an infinite

number of jumps in any finite time interval. However by a limiting argument, we can ignore the jumps of negligible size (say $< \epsilon$). Hence our space reduces to

$$\Gamma_\epsilon = (-\epsilon, \epsilon)^c \times [0, 1].$$

Given the jumps sizes $\{w_j\}$, jump locations $\{u_j\}$, and the number of jumps J, the prior probability density function (2.17) is

$$\pi(Z) = \Big[\Pi_{j=1}^{J} |w_j|\Big]^{1-\alpha} e^{2(\epsilon^{-1} - \epsilon^{-\alpha})} \alpha^J, \quad |w_j| \geq \epsilon \tag{2.20}$$

with respect to a Cauchy random field.

Using thus prior is essentially the same as using a penalty term in a regularization approach. For the SαS process, we have

$$\log \pi(Z) \propto J \log \alpha + (1 - \alpha) \Big(\sum_j \log |u_j| 1_{|u_j| > \epsilon} \Big) + \text{constant}. \tag{2.21}$$

The first term is an AIC like penalty for the number of knots J and the second term is a LASSO-type penalty in log-scale. There is also a hidden penalty which shrinks all the coefficients with magnitude less than ϵ to zero.

2.4.4 Computational and modeling considerations

The computational and modeling issues involved in choosing process priors, especially in high dimensional settings, are at the heart of non-parametric Bayesian modeling. In this section we discuss these issues for the models discussed in the previous section.

A main challenge with Gaussian process models is that a finite dimensional representation of the sample path is required for computation. For low dimensional problems (say $d \leq 3$), a reasonable approach is to place a grid on $\mathcal{X}$. Then we can approximate a continuous process Z by its values on the finitely many points $\{u_j\}_{j=1}^m$

on the grid. Using this approximation, our kernel model (2.10) can be written as

$$f(x) = \sum_{j=1}^{m} w_j K(x, u_j),$$

and the implied prior distribution on $(w_1, \ldots, w_m)$ is a multivariate normal with mean and covariance structure as defined by the kernel K evaluated at points $\{u_j\}$. For low-dimensional data a grid can be placed on the input space. However, this approach is not practical in higher dimensions. This issue is addressed in Gaussian process regression models by evaluating the function at the training and future test data points. This corresponds to a fixed design setting. It is important to note however, that the prior being sampled in this model is not over $\mathcal{X}$ but the restriction of $\mathcal{X}$ to the data. Both the direct model and the kernel model will face this computational consideration and thus the computational cost will not differ significantly between models.

For pure jump processes discretization is not the bottleneck. The nature of the pure jump process ensures that the kernel model will have discrete knots. The key issue in using a pure jump processes to model multivariate data is that the knots of the model should be representative of samples drawn from the marginal distribution of the data ρ_X. This is a serious computational as well as modeling challenge, it is obvious that independently sampling each dimension will typically not be a good idea either in terms of computational time or modeling accuracy. In Section 2.5.2 we provide a kernel model that addresses this issue.

A theoretical and empirical comparison of the accuracy of the various process priors on a variety of function classes and data sets would be of interest, but is beyond the scope of this paper. Due to the extensive literature on Gaussian process models from theoretical as well as practical perspectives [Rasmussen and Williams 2006, Ghosal and Roy 2006] our simulations will focus on two pure jump process

models.

2.5 Posterior Inference

For the case of regression our model is

$$y_i = f(x_i) + \varepsilon_i \qquad \text{for } x_i \in \mathcal{X}$$

with $\{\varepsilon_i\}$ as normal independent random variables and the unknown regression function f (which is assumed to be in $\mathcal{H}_K$) is modeled as

$$f(x) = \int_{\mathcal{X}} K(x,u)Z(du).$$

In the case of binary regression we can use a probit model

$$\mathbb{P}(y_i = 1|x_i) = \Phi[f(x_i)], \tag{2.22}$$

where $\Phi[\cdot]$ is the cumulative distribution function of the standard normal distribution.

In Section 2.4, we discussed specifying a prior on $\mathcal{H}_K$ via the random measure $Z(du)$. The observed data add to our knowledge of both the "true function" $f(\cdot)$ and the distribution of $Z(du)$. This information is used to update the prior and obtain the posterior density $\pi(Z|D)$. For pure jump measures $Z(du)$ and most nonparametric models this update is computationally difficult because there is no closed-form expression for the posterior distribution. However, Markov chain Monte Carlo (MCMC) methods can be used to simulate the posterior distribution.

We will apply a Dirichlet process model to a high-dimensional binary regression problem and illustrate the use of Lévy process models on a univariate regression problem.

2.5.1 Lévy process model

Posterior inference for Lévy random measures have been less explored than Dirichlet and Gaussian processes. Wolpert *et al.* [2003] is a recent comprehensive reference on this topic. We use the methodology developed in this work for our model.

The random measure $Z(du)$ is given by

$$Z(du) \sim \text{Lévy}(\nu(dw)du)$$

where

$$\nu(dw) = \frac{\Gamma(\alpha+1)}{\pi} \sin\left(\tfrac{\pi\alpha}{2}\right) |w|^{-1-\alpha}\, 1_{\{w:|w|>\epsilon\}}\, dw \qquad \alpha \in (0,2]$$

is the Lévy measure (truncated) for the SαS process. As explained in Section 2.4.3, since $\nu(dw)$ is not a finite measure on $\mathbb{R}$, we ignore jumps of size smaller than ϵ. Any realization of the random measure $Z(du)$ is an element of the parameter space Θ

$$\Theta := \bigcup_{J=0}^{\infty} \Big((-\epsilon, \epsilon)^c \times [0,1] \Big)^J$$

with the prior probability density function given by Equation(2.20), with respect to a Cauchy random field.

Transition probability proposal

In this section, we describe an MCMC algorithm to simulate from Θ according to the posterior distribution. We construct an irreducible transition probability distribution $Q(d\theta^*|\theta)$ on the parameter space Θ such that the stationary distribution of the chain will be the posterior distribution.

Two different realizations from the parameter space Θ may not have the same number of jumps. Hence the number of jumps J is modeled a birth-death process. At any iteration step t the parameter space consists of J jump locations $\{u_j\}$ of size

$\{w_j\}$, $\theta_t = \{w_j, u_j\}_{j=1}^{J}$. The (weighted) transition probability algorithm, Algorithm 1, computes the weighted transition probability to a new state θ^* given the current state θ.

Algorithm 1: Weighted transition probability algorithm $\mathcal{Q}(\theta)$.

input : $0 < p_b, p_d < 1$, $\tau > 0$, current state $\theta \in \Theta$

return: proposed new state θ^* and its weighted transition probability $Q(\theta^*|\theta)\pi(\theta)$

Draw $t \sim U[0,1]$;
if $t < 1 - p_b$ **then**
  draw uniformly $j \in \{1, ..., J\}$; draw $\gamma_1, \gamma_2 \sim \mathbf{No}(0, \tau^2)$;
  $w_* \leftarrow w_j + \gamma_1$; $u_* \leftarrow u_j + \gamma_2$;
  if $(|w_*| < \epsilon$ ***or*** $t < p_d)$ **then**
    $J \leftarrow J - 1$; delete (w_j, u_j);
    $Q(\theta^*|\theta)\pi(\theta) \leftarrow \frac{(J+1)p_b}{2\epsilon^{-\alpha}\left((1-p_b-p_d)\left[\Phi(\frac{w_j+\epsilon}{\tau}) - \Phi(\frac{w_j-\epsilon}{\tau})\right] + p_d\right)}$;
  else
    $Q(\theta^*|\theta)\pi(\theta) \leftarrow \left|\frac{w_*}{w_j}\right|$; $w_j \leftarrow w_*$; $u_j \leftarrow u_*$;
else
  $J \leftarrow J + 1$; $u_{_J} \sim U[\mathcal{X}]$; $w_{_J} \sim$ Birth;
  $Q(\theta^*|\theta)\pi(\theta) \leftarrow \frac{2\epsilon^{-\alpha}\left((1-p_d-p_b)\left[\Phi(\frac{w_J+\epsilon}{\tau}) - \Phi(\frac{w_J-\epsilon}{\tau})\right] + p_d\right)}{p_b J}$;

In the above algorithm, $\mathbf{No}(0, \tau^2)$ denotes the normal distribution with mean 0 and variance τ^2 and $\Phi(\cdot)$ denotes the distribution function of the standard normal distribution. The variables (p_b, p_d) stand for probability of birth step and death step respectively. There is an implicit update step, where a chosen point(u_j) is 'updated' with another point(u_*) with probability $1 - p_b - p_d$. In the birth step, a new point is sampled according to the density

$$\frac{\alpha|w|^{-1-\alpha}}{2\epsilon^{-\alpha}} \quad \epsilon > 0.$$

The MCMC algorithm

The MCMC algorithm, Algorithm 2, simulates draws from the posterior distribution. This is done by Metropolis-Hastings sampling using the weighted transition probability algorithm above to generate a Markov chain whose equilibrium density is the posterior density.

Algorithm 2: MCMC algorithm

input : data D, number of iterations T, weighted transition probability algorithm $\mathcal{Q}(\theta)$

return: parameters drawn from the posterior $\{\theta_i\}_{i=1}^{T}$

$J \sim \mathbf{Po}(2\epsilon^{-\alpha})$; `// initialize J`

for $j \leftarrow 1$ **to** J **do**

  `// initialize` $\theta(0)$

  $u_j \sim U[\mathcal{X}]$; $w_j \sim$ Birth;

for $t \leftarrow 1$ **to** T **do**

  `//` t`-th iteration of the Markov chain`

  $\{\theta_*, Q(\theta_*|\theta_t)\pi(\theta_t)\} \leftarrow \mathcal{Q}(\theta(t))$; `// call the weighted transition probability algorithm`

  $\log \pi(\theta_*|D) - \log \pi(\theta_t|D) = \log \frac{L(D|\theta_*)}{L(D|\theta_t)} + \log \frac{\pi(\theta_*)}{\pi(\theta_t)}$;

  $\zeta_* \leftarrow \log \pi(\theta_*|D) + \log Q(\theta_t|\theta_*) - \log \pi(\theta_t|D) - \log Q(\theta_*|\theta_t)$; `// the Metropolis-Hastings log acceptance probability`

  $e \sim \mathbf{Ex}(1)$;

  if $e + \zeta_{t+1} > 0$ **then** $\theta_{t+1} \leftarrow \theta_*$ **else** $\theta_{t+1} \leftarrow \theta_t$;

The MCMC algorithm will provide us with T realizations of the jump parameters $\{\theta_t\}_{t=1}^{T}$. We assume that the chain reaches its stationary distribution after b iterations ($b \ll T$). For each of the $T - b$ realizations, we have a corresponding function

$$\hat{f}_t(x) = \sum_{i=1}^{J_t} w_{it} K(x, u_{it}),$$

where for the t-th realization J_t is the number of jumps, w_{it} is the magnitude of the i-th jump, and u_{it} is the position of the i-th jump. Point estimates can be made

by averaging $\hat{f}$ and credible intervals can be computed from the distribution of $\hat{f}$ to provide an estimate of uncertainty.

Illustration on simulated data

Data is generated from a noisy sinusoid

$$f(x_i) = \sin(2\pi x_i) + \varepsilon_i \quad \text{for} \quad x \in [0,1], \tag{2.23}$$

with $\varepsilon_i \overset{iid}{\sim} \mathbf{No}(0, .01)$, $\{x_i\}_{i=1}^{100}$ points equally spaced in $[0,1]$, and $\{y_i\}_{i=1}^{100}$ are computed by equation (2.23). We applied the SαS model with $\alpha = 1.5$ and a Gaussian kernel $K(x,u) = \exp\{(x-u)^2\}$ to this data. We set $\epsilon = 0.01$ and $(p_b, p_u, p_d) = (0.4, 0.2, 0.4)$, in algorithms 1 and 2. In Figure 2.1a-d we plot the target sinusoid, the function realized at an iteration t of the Markov chain, and the jump locations and magnitudes of the random measure. In Figure 2.1e,f we provide a plot of the target function, realization of the data, and the 95% point-wise credible band – the 95% credible interval at each point x_i.

2.5.2 Classification of gene expression data

For Dirichlet processes there is extensive literature on exact posterior inference using MCMC methods [West 1992, Escobar and West 1995, MacEachern and Müller 1998, Müller *et al.* 2004] as well as work on approximate inference using variational methods [Blei and Jordan 2006]. Recently Dirichlet process priors have been applied to a Bayesian kernel model for high dimensional data. For example in Liang *et al.* [2007b] and Liang *et al.* [2007a] the Bayesian kernel model was used to classify gene expression data as well as digits, the MNIST database. We apply this model to gene expression data consisting of microarray gene expression profiles from 190 cancer samples and 90 normal samples [Ramaswamy *et al.* 2001, Mukherjee *et al.* 2003], over 16,000 genes.

The model is based upon the integral operator given in equation (2.19)

$$f(x) = \int_{\mathcal{X}} K(x,u)\, Z(du) = \int_{\mathcal{X}} w(u)\, K(x,u)\, F(du),$$

where the random signed measure $Z(du)$ is modeled by a random probability distribution function $F(du)$ and a random weight function $w(u)$. We assume that the support of $Z(du)$ and $w(u)F(du)$ are equal. A key point in our model will be that if our estimate of F is discrete and puts masses w_i at support points (or "knots") u_i, then the expression for $f(\cdot)$ is simply

$$f(x) = \sum_i w(u_i)\, K(x,u_i). \tag{2.24}$$

The above model, in which basis functions are placed at random locations and a joint distribution is specified for the coefficients, has been considered previously in the literature (see Neal, R. M. [1996], Liang *et al.* [2007a]). In Liang *et al.* [2007a] uncertainty about F is expressed using a Dirichlet process prior, $\mathrm{Dir}(\alpha, F_0)$. The posterior after marginalization is also a Dirichlet distribution and given data $(x_1, \ldots, x_n)$ the posterior will have the following representation [Liang *et al.* 2007a,b]

$$\hat{f}(x) = \frac{\alpha}{\alpha+n} \int w(u)K(x,u)F_0(du) + \frac{1}{\alpha+n} \sum_{i=1}^{n} w(x_i)K(x,x_i),$$

which can be approximated by the following discrete summation

$$\hat{f}(x) \approx \sum_{i=1}^{n} w_i K(x,x_i) \tag{2.25}$$

when $\frac{\alpha}{n}$ is small and $w_i = \frac{w(x_i)}{\alpha+n}$. We specify a mixture-normal prior on the coefficients w_i as in Liang *et al.* [2007a] and use the same MCMC algorithm to simulate the posterior.

Note that although equation (2.25) has the same form as the representer theorem, it is derived from a very different formulation. In fact, when there is unlabeled data available – $(x_{n+1}, \ldots, x_{n+m})$ drawn from the margin ρ_X – our model has the following discrete representation

$$\hat{f}(x) = \sum_{i=1}^{n} w_i K(x, x_i) + \sum_{i=1}^{m} w_{i+n} K(x, x_{i+n}),$$

where $w_\ell = \frac{w(x_\ell)}{\alpha+m+n}$. The above form is identical to the one obtained via the manifold regularization framework [Belkin and Niyogi 2004, Belkin *et al.* 2006]. The two derivations are from different perspectives. This simple incorporation of unlabeled data into the model further illustrates the advantage of placing the prior over random measures in the Bayesian kernel model.

In our experiments we first applied a standard variation filter to reduce the number of genes to $p = 2800$. We then randomly assigned 20% of the samples from the cancer and normal groups to training data and use the remaining 80% as test data. We used a linear kernel in the model and we used the classification model detailed in [Liang *et al.* 2007a].

We performed two analyses on this data:

Analysis I – The training data were used in the model and the posterior probability was simulated for each point in the test set. A linear kernel was used.

Analysis II – The training and unlabeled test data were used in the model and the posterior probability was simulated for each point in the test set. A linear kernel was used.

The classification accuracy for Analyses I and II were 73% and 85%, respectively. The accuracy of the predictive models in Analysis I is comparable to that obtained for

support vector machines in Mukherjee *et al.* [2003]. Figure 2.2 displays boxplots of the posterior mean of the 72 the normal and 152 cancer samples for the two analyses.

2.6 Discussion

The modeling objective underlying this paper is to formulate a coherent Bayesian perspective for regression using a RHKS model. This requires a firm theoretical foundation characterizing the function space that the Bayesian kernel model spans and the relation of this space to the RKHS. Our results in Section 2.2 are interesting in their own right, in addition to providing this foundation.

We examined the function class defined by the Bayesian kernel model, the integral of a kernel with respect to a signed Borel measure

$$\mathcal{G} = \left\{ f \;\middle|\; f(x) = \int_{\mathcal{X}} K(x,u)\ \gamma(\mathrm{d}u), \quad \gamma \in \Gamma \right\}, \tag{2.26}$$

where $\Gamma \subseteq \mathscr{B}(\mathcal{X})$. We stated an equivalence under certain conditions of the function class $\mathcal{G}$ and the RKHS induced by the kernel. This implies: (a) a theoretical foundation for the use of Gaussian processes, Dirichlet processes, and other jump processes for non-parametric Bayesian kernel models, (b) an equivalence between regularization approaches and the Bayesian kernel approach, and (c) an illustration of why placing a prior on the distribution is natural approach in Bayesian non-parametric modelling.

Coherent non-parametric methods have been of great interest in the Bayesian community, however function analytic issues have not been considered. Conversely theoretical studies of RKHS have not approached the approximation and estimation problems from a Bayesian perspective (the exception to both of these are the works of Wahba [1990] and Diaconis [1988]). It is our view that the interface of these perspectives is a promising area of research for statisticians, computer scientists, and mathematicians and has both theoretical and practical implications.

A better understanding of this interface may lead to a better understanding of the following research problems:

1. Posterior consistency: It is natural to expect the posterior distribution to concentrate around the true function since the posterior distribution is a probability measure on the RKHS. A natural idea is to use the equivalence between the RKHS and our Bayesian model to exploit the well understood theory of RKHS in proving posterior consistency of the Bayesian kernel model. Tools such as concentration inequalities, uniform Glivenko-Cantelli classes, and uniform central limit theorems may be helpful.

2. Priors on function spaces: In this paper we discuss general function classes without concern for more subtle smoothness properties. An obvious question is can we use the same ideas to relate priors on measures and the kernel to specific classes of functions, such as Sobolev spaces. A study of the relation between integral operators and priors could lead to interesting and useful results for putting priors over specific function classes using the kernel model.

3. Comparison of process priors for modeling: A theoretical and empirical comparison of the accuracy of the various process priors on a variety of function classes and data sets would be of great practical importance and interest, especially for high dimensional problems.

4. Numerical stability and robust estimation: The original motivation for regularization methods was to provide numerical stability in solving Fredholm integral equation of the first kind. Our interest is that of providing robust nonparametric statistical estimates. A link between stability of operators and the generalization or predictive ability of regression estimates is known [Bousquet and Elisseeff 2002, Poggio *et al.* 2004]. Further developing this relation is a

very interesting area of research and may be of importance for the posterior consistency of the Bayesian kernel model.

5. The results we have presented extend to the multivariate case (see theorem 2.6.1) without complication. The simplest multivariate extension is to assume independence of the dimensions, however for small sample sizes and many dimensions this is not practical. This issue can be addressed by carefully inducing covariance structure in the model [Liang *et al.* 2007a,b].

   Here we give the statement of the multivariate version of the Lévy-Khintchine formula [Applebaum 2004, Corollary 2.4.20].

   Theorem 2.6.1. *(**Lévy-Khintchine**) Let X be a d-dimensional Lévy process with characteristic function $\phi_t(u) := \mathbb{E}(e^{i\langle u, X_t\rangle}), u \in \mathbb{R}^d$. Then there exists a unique vector $a \in \mathbb{R}^d$, a $d \times d$ semi-positive definite matrix σ, and ν a positive measure on $\mathbb{R}^d \backslash 0$ with $\int_{\mathbb{R}^d}(1 \wedge |u|^2)\nu(du) < \infty$ such that,*

$$\phi_t(u) = \exp\left\{t\left[i\langle u, a\rangle - \frac{1}{2}\langle u, \sigma u\rangle + \int_{\mathbb{R}^d\backslash 0}[e^{i\langle u,s\rangle} - 1 - i\langle u, s\rangle 1_{\{s:|s|<1\}}(s)]\nu(ds)\right]\right\}$$

   where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^d$.

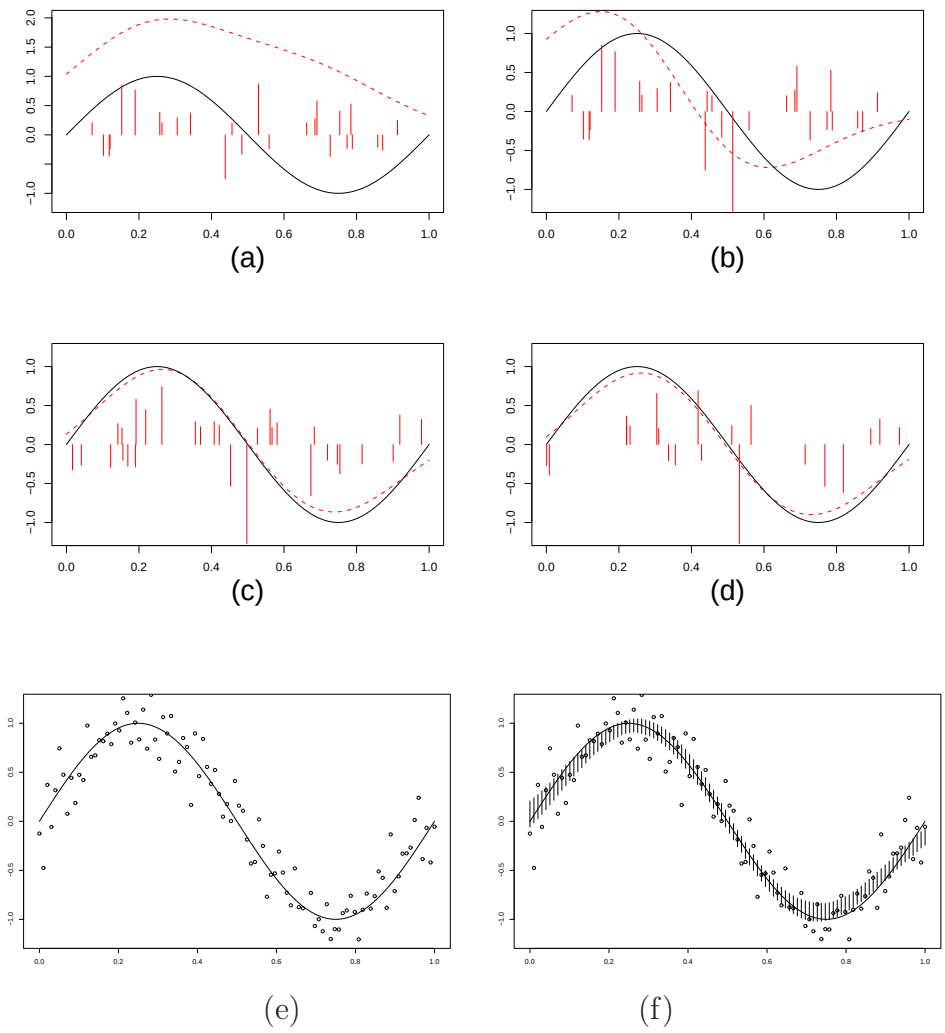

Figure 2.1: Plots of the target sinusoid (solid line), the function realized at an iteration t of the Markov chain (dashed line), and the jump locations and magnitudes of the measure (spikes) for (a) $t = 1$, (b) $t = 10$, (c) $t = 5 \times 10^3$, and (d) $t = 10^4$. (e) A realization of the simulated data (circles) and the underlying target sinusoid (solid line). (f) The 95% point-wise credible band for the data and the target sinusoid.

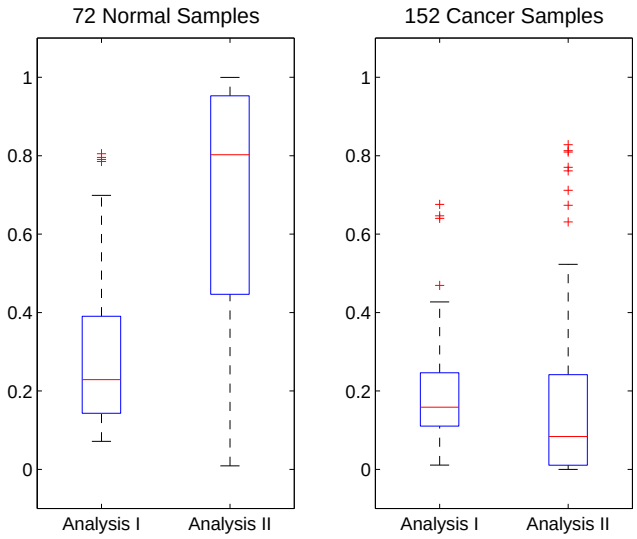

Figure 2.2: Boxplots of the posterior mean for normal and cancer samples with just the training data (Analysis I) and the training and unlabeled test data (Analysis II). (In the above boxplots, the box ranges from the first quartile (F.Q.) to the third quartile (T.Q.) of the data, while the line shows the median. The dots denote the outliers, which are points which lie beyond 1.5*(T.Q. - F.Q.) on either side of the box.)

Chapter 3

Posterior Consistency

3.1 Introduction

In this paper we study the posterior consistency of certain Bayesian nonparametric regression models using Lévy random measures as priors. Nonparametric function estimation using Lévy random measures has received great attention in the literature during the last decade. They were introduced in [Wolpert. and Ickstadt 1998, Wolpert and Ickstadt 1998] in which the authors modeled spatial data using Lévy random fields. Other applications can be found in [Wolpert *et al.* 2003, Wolpert and Ickstadt 2004], also in which efficient numerical schemes for implementation are developed.

A similar approach is taken in Nieto-Barajas *et al.* [2004], in which normalized random measures derived from increasing additive processes are considered as priors for distribution functions. The main difference in our approach is that we consider random measures which are not necessarily positive. Moroever in our case, the random measures can be unbounded and need not even have finite total variation.

Recently [Tu *et al.* 2006] extended the methodology developed in the papers mentioned above to construct a richer class of models, called LARK (Lévy Adaptive

Regression Kernels), for function estimation. Consider estimating an unknown function based on noisy data. The traditional approach is to represent the function in a series expansion using a linear combination of basis functions and then to estimate the coefficients from the data. In [Tu *et al.* 2006], the authors use overcomplete dictionaries instead of an orthonormal basis. Over complete representation may lead to possibly non-unique coefficients, but facilitates sparser representations by using fewer non-zero coefficients. Lévy random fields are natural candidates for constructing prior distributions on functions using these overcomplete representations. In most situations closed form expressions for posterior distributions are not available, but powerful algorithms exist to sample from the posterior distribution. See Clyde and Wolpert [2007] for a thorough exposition of the relative merits of these models and a comparison with other relevant models in the literature.

Even though these models perform very well in practice, the issue of their posterior consistency has not been studied in the literature. Posterior consistency of infinite dimensional non parametric models is not always guaranteed and the modeler needs to be careful to avoid unpleasant surprises (see Diaconis and Freedman [1986a,b], Barron *et al.* [1999]). One of the important criteria for the posterior consistency to hold is that the prior puts positive mass on every Kullback-Liebler (KL) neighborhoods. This condition is easy to verify when the number of dictionary elements is almost surely finite. However, in the models we consider it is possible to have an infinite number of dictionary elements in the representation of the unknown function, and more importantly the coefficients corresponding to those elements need not be absolutely summable. Hence it requires more effort to verify the KL positivity condition mentioned above, and we accompolish this using a limiting arguement.

In addition to verifying the KL positivity condition, one has to show the existence of tests which separate points in the parameter space. This depends on the the

specific model under consideration. Posterior consistency for independent but not identically distributed random variables are studied only recently in the literature. The general framework for proving such results was laid out by Amewou-Atisso *et al.* [2003], Ghosal and van der Vaart [2007a,b] and Choi and Schervish [2007]. Choi and Schervish [2007] prove the posterior consistency for a regression model with Gaussian error terms. Recently posterior consistency was proved for nonparametric binary regression by Ghosal and Roy [2006], who conjectured that similar results might hold for the Poisson regression. We verify that conjecture, proving posterior consistency for the Poisson regression model under moderate assumptions. Our results on the Poisson regression model are of independent interest.

The rest of the paper is organized as follows. Section 2 explains the basic frame work of Lévy random measures. Section 3 contains the main results of the paper. Section 4 has examples which use the results from the previous sections. Section 5 contains a brief discussion and conclusion.

3.2 Kernel Model

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}$ and Θ be a space of real valued functions on $\mathcal{X}$. Let Ω be a complete separable metric space and $K : \mathcal{X} \times \Omega \to \mathbb{R}_+$ a Borel measurable function. Set

$$\Theta_D \equiv \left\{ \theta \in \Theta : \theta(x) = \sum_{j=1}^{J} u_j K(x, \omega_j), \qquad J \in \mathbb{N},\ u_j \in \mathbb{R},\ \omega_j \in \Omega \right\}. \tag{3.1}$$

The function $K(\cdot, \cdot)$ will be chosen later to ensure that Θ_D is dense in Θ, in some specified topology. Let $M_D(\Omega)$ denote the space of finitely supported discrete signed measures on Ω. Let $\mathcal{K} : M_D(\Omega) \mapsto \Theta_D$ be the integral operator defined by

$$\mathcal{K}\big[\mathcal{L}\big](x) \equiv \int_{\Omega} K(x, \omega)\mathcal{L}(d\omega), \quad x \in \mathcal{X} \tag{3.2}$$

where $\mathcal{L} \in M_D(\Omega)$ is given by

$$\mathcal{L}(d\omega) \equiv \sum_{j=1}^{J} u_j \delta_{\omega_j}(d\omega), \quad J \in \mathbb{N}, \ \{u_j, \omega_j\} \in \mathbb{R} \times \Omega. \tag{3.3}$$

The integral operator $\mathcal{K}$ will be extended later to countably supported measures $\mathcal{L}$ on Ω with possibly infinite total variation. We define a probability distribution Π on the measurable space $\left(\Theta_D^-, \mathcal{B}(\Theta_D^-)\right)$ (here Θ_D^- denotes the closure of Θ_D in some topology to be defined later) by first specifying a joint distribution of $\left(\mathcal{L}(A_1), \mathcal{L}(A_2), \ldots, \mathcal{L}(A_n)\right)$, for disjoint Borel sets $A_i \subset \Omega$, and define Π to be the corresponding probability measure induced on $\left(\Theta_D^-, \mathcal{B}(\Theta_D^-)\right)$ by the integral operator $\mathcal{K}$.

The joint distribution of $\left(\mathcal{L}(A_1), \mathcal{L}(A_2), \ldots, \mathcal{L}(A_n)\right)$ is determined by the joint distribution of the random variables J and $\{u_j, \omega_j\}$. Next, we introduce Lévy random fields for defining a joint distribution for these random variables [Tu *et al.* 2006].

3.2.1 Lévy Random Measures

Let $\pi(du\, d\omega)$ be a probability measure on $\mathbb{R} \times \Omega$ and $\nu_+ > 0$. Let $J \sim \text{Po}(\nu_+)$ and, conditional on J, let $\{u_j, \omega_j\}_{j=1}^{J} \overset{iid}{\sim} \pi(du\, d\omega)$. The random measure $\mathcal{L}$ defined by

$$\mathcal{L}(A) = \sum_{j=1}^{J} 1_A(\omega_j)\ u_j \tag{3.4}$$

assigns independent infinitely-divisible (ID) random variables $\mathcal{L}(A_i)$ to disjoint Borel sets $A_i \subset \Omega$. For any Borel set $A \subset \Omega$, $\mathcal{L}(A)$ has characteristic function

$$\mathbb{E}\left[e^{it\mathcal{L}(A)}\right] = \exp\left\{\iint_{\mathbb{R}\times A} (e^{itu} - 1)\nu(du\, d\omega)\right\} \tag{3.5}$$

where $\nu(du\, d\omega) \equiv \nu_+ \pi(du\, d\omega)$. A sigma-finite random measure $\mathcal{L}$ with characteristic function given by (3.5) can be well defined for any sigma-finite positive measure

$\nu(du\,d\omega)$ on $\mathbb{R}\times\Omega$ that satisfies

$$\iint_{\mathbb{R}\times K}(1\wedge|u|)\ \nu(du\,dw)<\infty \tag{3.6}$$

for each compact $K\subset\Omega$. If a measure $\nu(\cdot)$ satisfying (3.6) is infinite, then the random measure $\mathcal{L}$ will have a countably-infinite support set. For relatively compact Borel sets A (*i.e.*, those with compact closure), the sum

$$\mathcal{L}(A)=\sum 1_A(\omega_j)u_j \tag{3.7}$$

converges absolutely.

More generally, for any positive sigma-finite measure $\nu(du\,d\omega)$ on $\mathbb{R}\times\Omega$ satisfying the weaker condition

$$\iint_{\mathbb{R}\times K}(1\wedge u^2)\ \nu(du\,d\omega)<\infty \tag{3.8}$$

for each compact $K\subset\Omega$, and any sigma-finite signed measure $m(d\omega)$, there exists a random sigma-finite measure $\mathcal{L}(d\omega)$ with characteristic function

$$\mathbb{E}\left[e^{it\mathcal{L}(A)}\right]=\exp\left\{itm(A)+\iint_{\mathbb{R}\times A}\big(e^{itu}-1-ith_0(u)\big)\nu(du\,d\omega)\right\} \tag{3.9}$$

with "compensator function" $h_0(u)\equiv u1_{|u|\le1}(u)$ (see Rajput and Rosiński [1989, Prop 2.1] for extending the classic Lévy Khintchine formula to random measures). Since $e^{itu}-1-ith_0(u)=O(u^2),\ u\to0$, the integral in (3.9) converges for any relatively compact A. Equation (3.9) remains valid, if we replace $h_0(u)$ by any other bounded $h(u)$ satisfying $h(u)=u+O(u^2),\ u\to0$, and $m(A)$ by $m(A)+\iint_{\mathbb{R}\times A}[h(u)-h_0(u)]\nu(du\,d\omega)$.

3.2.2 Lévy Random Fields

In this section we construct Lévy random fields. Let $\nu(du\,d\omega)$ be a positive sigma-finite measure on $\mathbb{R}\times\Omega$ satisfying the integrability condition (3.8). Let

$$N(du\,d\omega)\sim \mathrm{Po}(\nu) \tag{3.10}$$

be a Poisson random measure on $\mathbb{R}\times\Omega$ which assigns independent $\mathrm{Po}\big(\nu(B_i)\big)$ distributions to disjoint Borel sets $B_i\subset\mathbb{R}\times\Omega$. Let

$$\tilde{N}(du\,d\omega)\equiv N(du\,d\omega)-\nu(du\,d\omega) \tag{3.11}$$

denote the *compensated* Poisson measure with mean 0, an isometry from $L_2\big(\mathbb{R}\times\Omega,\nu(du\,d\omega)\big)$ to the square-integrable zero-mean random variables [iti Sato 1999, *pg.* 38]. For relatively compact Borel sets $A\subset\Omega$ and a sigma-finite signed measure m, define (following Wolpert and Taqqu [2005])

$$\mathcal{L}(A)\equiv m(A)+\iint_{\mathbb{R}\times A}\big(u-h(u)\big)N(du\,d\omega)+\iint_{\mathbb{R}\times A}h(u)\tilde{N}(du\,d\omega) \tag{3.12}$$

where $h(u)$ is any bounded function satisfying $h(u)=u+O(u^2)$, $u\to 0$. The random variable $\mathcal{L}(A)$ has the characteristic function

$$\mathbb{E}\left[e^{it\mathcal{L}(A)}\right]=\exp\left\{itm(A)+\iint_{\mathbb{R}\times A}\big(e^{itu}-1-ith(u)\big)\nu(du\,d\omega)\right\}. \tag{3.13}$$

Henceforth we denote the Lévy random measure $\mathcal{L}$ with intensity measure ν

$$\mathcal{L}\sim \text{Lévy}\big(\nu,m,h\big). \tag{3.14}$$

In what follows, unless otherwise specified, we set $m\equiv 0$ and $h(u)=h_0(u)\equiv u1_{|u|<1}$. If in addition, the measure ν satisfies the stronger integrability condition (3.6), we also set $h\equiv 0$. In either of these instances, we further omit m,h from the notation and write

$$\mathcal{L}\sim \text{Lévy}\big(\nu\big). \tag{3.15}$$

Notice that if $\nu_+ \equiv \nu(\mathbb{R}\times\Omega) < \infty$, then the random measure $\mathcal{L}$ can be represented as in (3.4). A similar representation is available (with $J = \infty$) if ν is infinite but satisfies (3.6), but no such convergent sum is available if (3.6) fails.

Following [Taqqu and Wolpert 1983, section 3], we now extend the definition of the integral operator $\mathcal{K}$ as follows: For the random measure $\mathcal{L} \sim \text{Lévy}\big(\nu(du\,d\omega)\big)$, set

$$\mathcal{K}\big[\mathcal{L}\big](x) \equiv \int_\Omega K(x,\omega)\mathcal{L}(d\omega),\ x\in\mathcal{X} \tag{3.16}$$

$$\equiv \iint_{\mathbb{R}\times\Omega} K(x,\omega)\big(u-h(u)\big)N(du\,d\omega) + \iint_{\mathbb{R}\times\Omega} K(x,\omega)h(u)\tilde{N}(du\,d\omega) \tag{3.17}$$

for all Borel measurable $K(\cdot,\cdot)$ for which the integrals in (3.17) converge – *i.e.,* for those in "Musielak-Orlicz" (Rajput and Rosiński [1989]) space for (3.17) to be well defined. For a function $f : \mathbb{R}\times\Omega \mapsto \mathbb{R}$ define

$$I(f) \equiv \iint_{\mathbb{R}\times\Omega} f(u,\omega)\tilde{N}(du\,d\omega). \tag{3.18}$$

Define

$$L_{\Psi_p}\big(\mathbb{R}\times\Omega,\nu\big) \equiv \left\{ f : f \text{ is Borel measurable and } \iint_{\mathbb{R}\times\Omega} \Psi_p\big(f(u,\omega)\big)\nu(du\,d\omega) < \infty \right\} \tag{3.19}$$

where

$$\Psi_p(y) \equiv \begin{cases} |y| \wedge |y|^2, & 0 \le p \le 1, \\ |y|^p \wedge |y|^2, & 1 < p \le 2, \\ |y|^p \vee |y|^2, & p > 2. \end{cases} \tag{3.20}$$

Theorem 3.2.1. *(Rajput and Rosiński [1989], Theorem 3.3, Gaigalas [2004a,b]) The integral $I(f)$ in equation (3.18) is well defined and $\mathbb{E}\big(|I(f)|^p\big) < \infty$ for some $p > 0$, if and only if $f \in L_{\Psi_p}\big(\mathbb{R}\times\Omega,\nu\big)$.*

Remark 3.2.2. *For any $A \subset \Omega$ with compact closure, $f(u,\omega) \equiv h_0(u)\,1_A(\omega) \in L_{\Psi_0}\big(\mathbb{R}\times\Omega,\nu\big)$, since*

$$\iint_{\mathbb{R}\times A}\big(|f(u,\omega)|\wedge|f(u,\omega)|^2\big)\nu(du\,d\omega) = \iint_{[-1,1]\times A} u^2\nu(du\,d\omega) < \infty.$$

Hence the compensated random measure $\mathcal{L}(A)$ given by (3.12) *is well defined. Moreover for $p \geq 1$,*

$$\mathbb{E}\big(\big|\mathcal{L}(A)\big|^p\big) < \infty, \text{ if and only if } \iint_{[-1,1]^c\times A}|u|^p\nu(du\,d\omega) < \infty. \tag{3.21}$$

Remark 3.2.3. *Also*

$$L_p\big(\mathbb{R}\times\Omega,\nu\big) \subset L_{\Psi_p}\big(\mathbb{R}\times\Omega,\nu\big),\ 1 \leq p \leq 2,$$

with strict inclusion unless $p = 2$.

3.2.3 Truncation

Let ν be an infinite Lévy measure and let $\mathcal{L} \sim \text{Lévy}\big(\nu\big)$. Then $\mathcal{L}$ has an infinite number of elements in its support almost surely. Hence we will require a sequence of Lévy random measures $\{\mathcal{L}^\epsilon\}$ with almost surely finite number of points in their support and approximate (made more precise below) $\mathcal{L}$ as ϵ goes to 0. This finite approximation is also useful and routinely implemented in numerical procedures to sample Lévy random measures. Hence it is useful to study the weak convergence of the sequence of random measures $\{\mathcal{L}^\epsilon\}$ to a limiting $M(\Omega)$ valued random variable $\mathcal{L}$, where $M(\Omega)$ denotes the set of finite signed measures on Ω.

If Ω is compact, by the Reisz representation theorem [Albiac and Kalton 2006, Theorem 4.1.1, pg. 74], $M(\Omega)$ can be identified as the dual of the Banach space $C(\Omega)$ of

continuous functions on Ω. The duality is given by

$$\mu(f) = \int_\Omega f d\mu, \ f \in C(\Omega), \ \mu \in M(\Omega)$$

with total variation norm $\|\mu\| \equiv |\mu|(\Omega)$. A natural topology on $M(\Omega)$ is the usual weak star topology on $M(\Omega)$ [Reed and Simon 1980, pg. 113], the weakest topology for which all the functions $\mu \mapsto \mu(f), \ f \in C(\Omega)$ are continuous. Finite linear combinations of point measures are dense in $M(\Omega)$ [Reed and Simon 1980, pg. 114], so $M(\Omega)$ is the closure of $M_D(\Omega)$ $\big($the set of finite discrete measures on $M(\Omega)\big)$ in the weak star topology. Thus it is natural to approximate countably supported Lévy random measures with a sequence of measures belonging to $M_D(\Omega)$. Weak convergence of probability measures is well studied in the metric space setting [Billingsley 1999]. Alas, since the Banach space $C(\Omega)$ is infinite dimensional, the weak star topology on its dual $M(\Omega)$ is not metrizable (see [Dunford and Schwartz 1988, pg. 462] & [Wehausen 1938, Theorem 17]).

Hence to study the weak convergence of random variables in $M(\Omega)$, one option is to consider other natural topologies on $M(\Omega)$ which *are* metrizable. Another option is to consider $M(\Omega)$ as a subset of $D(\Omega)$, the set of tempered distributions ([Reed and Simon 1980, pg. 134]) on Ω. Instead we settle for a simpler approach: We find a sequence of Lévy random measures $\{\mathcal{L}^\epsilon\}$ such that for any precompact $A \subset \Omega$, the random variables $\{\mathcal{L}^\epsilon(A)\}$ converge in distribution to $\mathcal{L}(A)$.

Let Ω^ϵ be an increasing sequence of compact sets (as $\epsilon \to 0$) with $\cup_{\epsilon>0}\Omega^\epsilon = \Omega$. Define

$$I^\epsilon(u,\omega) \equiv 1_{\{|u|>\epsilon, \omega \in \Omega^\epsilon\}}(u,\omega) \in L_1\big(\Omega, \nu\big). \tag{3.22}$$

and set

$$\nu^\epsilon(du\, d\omega) \equiv I^\epsilon(u,\omega)\nu(du\, d\omega). \tag{3.23}$$

Consider

$$\mathcal{L}^\epsilon \sim \text{Lévy}\big(\nu^\epsilon, 0, h_0\big). \tag{3.24}$$

By equation (3.8), $\nu^\epsilon(\mathbb{R}\times\Omega) = \nu\big((-\epsilon,\epsilon)^c \times \Omega^\epsilon\big) < \infty$ and $\mathcal{L}^\epsilon$ will only have almost surely finitely many points in its support. We call $\mathcal{L}^\epsilon$ the compound Poisson approximation of $\mathcal{L}$.

Theorem 3.2.4. *For any precompact $A \subset \Omega$, the sequence of random variables $\mathcal{L}^\epsilon(A) \Rightarrow \mathcal{L}(A)$, as $\epsilon \to 0$.*

Proof. By (3.8) the log Characteristic function

$$\iint_{\mathbb{R}\times A} \big(e^{itu} - 1 - ith(u)\big)\nu(du\, d\omega)$$

for $\mathcal{L}(A)$ is well defined. For $\epsilon > 0, t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[e^{it\mathcal{L}^\epsilon(A)}\right] &= \exp\left\{\iint_{\mathbb{R}\times A} \big(e^{itu} - 1 - ith_0(u)\big)\nu^\epsilon(du\, d\omega)\right\} \\
&= \exp\left\{\iint_{\mathbb{R}\times A} \big(e^{itu} - 1 - ith_0(u)\big)I^\epsilon(u,\omega)\nu(du\, d\omega)\right\}. \\
&\to \exp\left\{\iint_{\mathbb{R}\times A} \big(e^{itu} - 1 - ith_0(u)\big)\ \nu(du\, d\omega)\right\}.
\end{aligned}$$

by Lebesgue's dominated convergence theorem. ■

Remark 3.2.5. *Notice that for $\epsilon > 0$, the random variable $\mathcal{L}(A) - \mathcal{L}^\epsilon(A)$ is independent of $\mathcal{L}^\epsilon(A)$, and has the characteristic function*

$$\mathbb{E}\left(e^{it\Big(\mathcal{L}(A)-\mathcal{L}^\epsilon(A)\Big)}\right) = \exp\left\{\iint_{(-\epsilon,\epsilon)\times A} \big(e^{itu} - 1 - ith_0(u)\big)\nu(du\, d\omega)\right\}.$$

Example 3.2.6. *Let $\Omega = [0,1]$. Let $\mathbf{D}[0,1]$ be the Skorohod space of right continous functions with left limits, (see [Billingsley 1999, Chapter 3]) and $\mathcal{D}$ be the associated Borel sigma field.*

Let $\nu(du\, d\omega)$ satisfy equation (3.6)*. Set (see the Inverse Lévy Measure Algorithm described in Wolpert and Ickstadt [1998])*

$$\mathcal{L}(d\omega) \equiv \sum_{j=1}^{\infty} u_j \delta_{\omega_j}(\omega),\ \{u_j, \omega_j\} \in \mathbb{R} \times [0,1].$$

Then the process X_t on $[0,1]$ defined by

$$X_t \equiv \sum_{j=1}^{\infty} u_j 1_{\{\omega_j \leq t\}}$$

is infinitely divisible and the sample path $X \equiv \{X_t, t \in [0,1]\}$ is a random element of $\mathbf{D}[0,1]$. The truncated random measure $\mathcal{L}^\epsilon$ is given by

$$\mathcal{L}^\epsilon(d\omega) \equiv \sum_{j=1}^{\infty} u_j 1_{\{|u_j|>\epsilon\}} \delta_{\omega_j}(\omega),\ \{u_j, \omega_j\} \in \mathbb{R} \times [0,1]$$

and the process X_t^ϵ defined analogously, is the usual compound Poisson process. The random measures $\mathcal{L}$ and $\{\mathcal{L}^\epsilon\}$ induce distributions on the measure space $\big(\mathbf{D}[0,1], \mathcal{D}\big)$. By theorem (3.2.4) X_t^ϵ converges in distribution to X_t for each fixed $t > 0$. Hence the sequence $\{X_t^\epsilon,\ t \in [0,1]\}$ of $\mathbf{D}[0,1]$ variables converges in distribution to $\{X_t, t \in [0,1]\}$. This construction needs a little more modification, when ν does not satisfy (3.6) *as illustrated in the next example.*

This example also suggests another alternative to study weak convergence of Lévy random measures. The Skorohod topology is defined for functions on the positive real line and we believe that generalizing it to functions on Polish spaces (Ω in our case) will be a useful path to follow.

Example 3.2.7. *For the choice of the compensator $h(u) \equiv \sin(u)$, the one dimen-*

sional stable distribution $\boldsymbol{St}(\alpha, \beta, \gamma, \delta)$ *on the real line has the characteristic function*

$$\Phi_{\boldsymbol{St}}(t) = \begin{cases} \exp\Big[-\gamma|t|^{\alpha} + it\Big\{\delta - \beta\gamma \tan\frac{\pi\alpha}{2}\ (1 - |t|^{\alpha-1})\Big\}\Big], & \alpha \neq 1, \\ \exp\Big[-\gamma|t| + it\Big\{\delta - \frac{2\beta\gamma}{\pi}\log|t|\Big\}\Big], & \alpha = 1, \end{cases} \tag{3.25}$$

with $0 < \alpha \leq 2$, $-1 \leq \beta \leq 1$, $0 < \gamma < \infty$, $-\infty < \delta < \infty$, *with the Lévy measure*

$$\nu(du) = \begin{cases} \gamma\frac{\alpha}{\pi}\ \Gamma(\alpha)\ \sin(\pi\alpha/2)\ |u|^{-1-\alpha}\ (1 + \beta\,\mathrm{sgn}\ u)\ du, & \alpha \neq 1, \\ \frac{\gamma}{\pi}\ |u|^{-2}\ (1 + \beta\,\mathrm{sgn}\ u)\ du, & \alpha = 1. \end{cases} \tag{3.26}$$

The alpha stable Lévy measure satisfies

$$\int_{\mathbb{R}} (1 \wedge |u|)\ \nu(du)du < \infty, \tag{3.27}$$

only for $0 < \alpha < 1$, *so compensation is needed for alpha* ≥ 1 *including the skewed Cauchy distribution (*$\boldsymbol{St}$*(1,1,1,0)) with Lévy measure* ν

$$\nu(du) = \frac{1}{\pi}\ |u|^{-2} 1_{u>0}.$$

Let $(u_j, s_j) \in \mathbb{R}_+ \times [0,1]$ *be generated from a Poisson random field with intensity measure* $\nu(du)ds$. *Then for any fixed* $t \in (0,1]$, *the sequence of random variables defined by*

$$X_t^{\epsilon} \equiv \Big\{ \sum u_j : u_j > \epsilon,\ s_j \leq t \Big\} \tag{3.28}$$

diverges a.s. as $\epsilon \to 0$. *However if we compensate* X_t^{ϵ} *to obtain,*

$$Y_t^{\epsilon} \equiv X_t^{\epsilon} - t\,\delta^{\epsilon},\ \delta^{\epsilon} \equiv \int_{\epsilon}^{\infty} \sin(u)\nu(du) < \infty, \tag{3.29}$$

then the sequence of random variables Y_1^{ϵ} *converges almost surely to a* $\boldsymbol{St}$*(1,1,1,0) random variable* Y_1. *Moreover, the sequence of* $\mathbf{D}[0,1]$ *valued random variables*

$$Y^{\epsilon} \equiv \Big\{ Y_t^{\epsilon}, t \in [0,1] \Big\} \tag{3.30}$$

converges in distribution to a skewed Cauchy process in $\mathbf{D}[0,1]$*, as* $\epsilon \downarrow 0$*.*

Let $Y_1^{\epsilon,1}, Y_1^{\epsilon,2}$ *be two independent copies of the random variable* Y_1^{ϵ} *defined in* (3.29)*. Then the sequence of random processes*

$$\tilde{Y_t}^{\epsilon} \equiv Y_t^{\epsilon,1} - Y_t^{\epsilon,2} = X_t^{\epsilon,1} - X_t^{\epsilon,2}$$

converges almost surely to a Cauchy process $\Big(\boldsymbol{St}(1,0,2t,0)\Big)$*.*

Remark 3.2.8. *There exist methods in the literature, other than the truncation approach, for approximating Lévy random fields that have infinite Lévy measure with those that have finite Lévy measure.*

The Lévy measure for the Symmetric Alpha Stable (Sα*S) process on the unit interval is given by*

$$\nu_\alpha(du\, d\omega) = c_\alpha \gamma\alpha\ |u|^{-\alpha-1} du\, d\omega,\ \{u,\omega\} \in \mathbb{R}\times[0,1],\ \alpha \in (0,2), \gamma > 0$$

with $c_\alpha \equiv \frac{\Gamma(\alpha)}{\pi}\sin(\frac{\pi\alpha}{2})$ *and*

$$\nu_\alpha(\mathbb{R}\times[0,1]) = 2c_\alpha\gamma\alpha \int_0^\infty |u|^{-\alpha-1} du\, d\omega = \infty,\ \alpha \in (0,2).$$

It can be shown that the Theorem(3.2.4) is also applicable to the sequence $\mathcal{L}^{\epsilon} \sim$ *Lévy*$\big(\nu_\alpha^{'\epsilon}(du\, d\omega)\big)$ *with*

$$\nu_\alpha^{'\epsilon}(du\, d\omega) \equiv c_\alpha\gamma\alpha\ \Big(|u| \vee \epsilon\Big)^{-\alpha-1} du\, d\omega,\ \{u,\omega\} \in \mathbb{R}\times[0,1],\ \alpha \in (0,2), \gamma > 0.$$

In Ouyang et al. *[2008], the authors use the following approximation for* ν_α*:*

$$\tilde{\nu}_\alpha^{\epsilon}(du\, d\omega) \equiv c_\alpha\gamma\alpha\ \frac{1}{(\epsilon^2 + |u|^2)^{(\alpha+1)/2}}\ du\, d\omega,\ \{u,\omega\} \in \mathbb{R}\times[0,1],\ \alpha \in (0,2), \gamma > 0.$$

Notice that $\nu_\alpha^{'\epsilon}(\mathbb{R}\times[0,1]) < \infty$ *and* $\tilde{\nu}_\alpha^{\epsilon}(\mathbb{R}\times[0,1]) < \infty$*.*

3.3 Prior positivity

Let $\nu(du\,d\omega)$ be a positive Radon measure on Ω such that

$$\iint_{\mathbb{R}\times\Omega}(1\wedge u^2)\nu(du\,d\omega)<\infty \tag{3.31}$$

and set $\mathcal{L}\sim$ Lévy(ν). For $\epsilon>0$, let $\nu^\epsilon(du\,d\omega)$ denote the measure defined in (3.23) and let $\mathcal{L}^\epsilon\sim$ Lévy$\big(\nu^\epsilon,0,h_0\big)$. For any Kernel K such that $K(\cdot,\omega)$ is continuous on $\mathcal{X}$ for each fixed $\omega\in\Omega$, the integral operator $\mathcal{K}$ given by (3.17) maps $\mathcal{L}$ and $\mathcal{L}_\epsilon$ into C$(\mathcal{X})$. Let Π, $\{\Pi^\epsilon\}$ be probability distributions on $\big(C(\mathcal{X}),\mathcal{B}(C)\big)$ induced by the operator $\mathcal{K}$ via the random measures $\mathcal{L}$ and $\{\mathcal{L}^\epsilon\}$ respectively.

Denote the modulus of continuity of a real valued function $f:\mathcal{X}\mapsto\mathbb{R}$ by:

$$W_f(\delta)\equiv\sup_{|s-t|<\delta}\Big\{|f(s)-f(t)|,\ s,t\in\mathcal{X}\Big\} \tag{3.32}$$

$$W_K(\delta)\equiv\sup_{\omega\in\Omega}W_{K(\cdot,\omega)}(\delta). \tag{3.33}$$

Since the function W_K is non-decreasing in δ its right-continuous inverse $W_K^{\leftarrow}$ is well defined.

We now introduce two conditions, either of which (as we will see in theorem (3.3.2)) is sufficient to ensure the weak convergence of Π^ϵ.

Condition A: The family of functions $\Big\{K(\cdot,\omega),\ \omega\in\Omega\Big\}$ is uniformly bounded and equicontinuous and, in addition to (3.31), the Lévy measure satisfies

$$\iint_{\mathbb{R}\times\Omega}(1\wedge|u|)\nu(du\,d\omega)<\infty. \tag{3.34}$$

Condition B: There exists $r>0$ such that

$$\int_0^r\frac{1}{\sqrt{W_K^{\leftarrow}(a)}}\,da<\infty. \tag{3.35}$$

Remark 3.3.1. *If the function $K(\cdot,\cdot)$ is uniformly bounded and uniformly continuous on $\mathcal{X}\times\Omega$, then $\Big\{K(\cdot,\omega),\ \omega\in\Omega\Big\}$ is equicontinuous. Also, if $K(\cdot,\omega)$ is uniformly Hölder continuous for some $\alpha>1/2$, i.e.*

$$W_K(\delta)\le C\delta^{\alpha}, \tag{3.36}$$

for some $\alpha>1/2$ and $C>0$, then condition B is satisfied as well.

Theorem 3.3.2. *The sequence of probability measures $\{\Pi^{\epsilon}\}$ converges weakly to Π in $C(\mathcal{X})$ if either (A) or (B) holds.*

Proof. Set

$$f^{\epsilon}(x)\equiv\mathcal{K}[\mathcal{L}^{\epsilon}](x),\quad f(x)\equiv\mathcal{K}[\mathcal{L}](x)\quad x\in\mathcal{X}. \tag{3.37}$$

First we claim that each finite dimensional random vector of the form

$$\Big(f^{\epsilon}(x_1),f^{\epsilon}(x_2),\cdots,f^{\epsilon}(x_n)\Big),\ n\in\mathbb{N},\ \{x_k\}_{k=1}^n\in\mathcal{X},$$

converges in distribution to the random vector $\big(f(x_1),f(x_2),\cdots,f(x_n)\big)$ as $\epsilon\to0$. To see this, as in the proof of Theorem 3.2.4, use dominated convergence theorem to verify that

$$\lim_{\epsilon\to0}\mathbb{E}\left[e^{\sum_{k=1}^n ia_kf^{\epsilon}(x_k)}\right]=\mathbb{E}\left[e^{\sum_{k=1}^n ia_kf(x_k)}\right],\ n\in\mathbb{N},\ \{x_k\}_{k=1}^n\in\mathcal{X},\{a_k\}_{k=1}^n\in\mathbb{R}.$$

Hence the finite dimensional distributions of $\{\Pi^{\epsilon}\}$ converge to those of Π. To show that $\{\Pi^{\epsilon}\}$ converges weakly, by Prohorov's theorem ([Billingsley 1999], Thm 5.1, pg 59) it is enough to show the tightness of the sequence $\{\Pi^{\epsilon}\}$. This is equivalent to showing that ([Billingsley 1999],Thm 7.5), for every $\eta>0$,

$$\lim_{\delta\to0}\limsup_{\epsilon\to0}\mathbb{P}\big(W_{f^{\epsilon}}(\delta)>\eta\big)=0. \tag{3.38}$$

Under condition A,

$$W_{f^\epsilon}(\delta) \le \int_\Omega \sup_{|x-y|<\delta} |K(x,\omega) - K(y,\omega)| \, \big|\mathcal{L}\big|^\epsilon(d\omega) \le W_K(\delta) \, |\mathcal{L}|^\epsilon(\Omega), \tag{3.39}$$

where $|\mathcal{L}|$ denotes the total variation norm of the measure $\mathcal{L}$. Hence

$$\lim_{\delta\to 0} \limsup_{\epsilon\to 0} \mathbb{P}\big(W_{f^\epsilon}(\delta) > \eta\big) \le \lim_{\delta\to 0} \limsup_{\epsilon\to 0} \mathbb{P}\Big(W_K(\delta)|\mathcal{L}|^\epsilon(\Omega) > \eta\Big).$$

Theorem (3.2.4) and (3.34) guarantee that the sequence $\{|\mathcal{L}|^\epsilon(\Omega)\}$ converges in distribution to $|\mathcal{L}|(\Omega)$, and that $|\mathcal{L}|(\Omega) < \infty$ a.s. Therefore $\lim_{\delta\to 0} W_K(\delta)|\mathcal{L}|(\Omega) = 0$ a.s., so (3.38) holds, and the theorem is proved under condition A.

Now suppose condition B holds. From equations (3.17) and (3.37), for every $x \in \mathcal{X}$, the random variable $f^\epsilon(x)$ can be written as the sum of two independent random variables

$$f^\epsilon(x) = \sum_j K(x,\omega_j) u_j 1_{|u_j|>1} + \iint_{[-1,1]\times\Omega} K(x,\omega) u \tilde{N}^\epsilon(du\, d\omega) \tag{3.40}$$

where $\tilde{N}^\epsilon(du\, d\omega) \equiv N^\epsilon(du\, d\omega) - \nu^\epsilon(du\, d\omega)$ is the fully compensated Poisson random measure, with $N^\epsilon(du\, d\omega) \sim \mathrm{Po}(\nu^\epsilon)$. Therefore

$$\begin{aligned}
\sup_{|x-y|<\delta} |f^\epsilon(x) - f^\epsilon(y)| &\le \sup_{|x-y|<\delta} \sum_j \Big|K(x,\omega_j) - K(y,\omega_j)\Big| |u_j| 1_{|u_j|>1} \\
&\quad + \sup_{|x-y|<\delta} \Big| \iint_{[-1,1]\times\Omega} \Big(K(x,\omega) - K(y,\omega)\Big) u \, \tilde{N}^\epsilon(du\, d\omega)\Big| \\
&\le W_K(\delta)\Big(\sum_j |u_j| 1_{|u_j|>1}\Big) \\
&\quad + \sup_{|x-y|<\delta} \Big| \iint_{[-1,1]\times\Omega} \Big(K(x,\omega) - K(y,\omega)\Big) u \, \tilde{N}^\epsilon(du\, d\omega)\Big|.
\end{aligned} \tag{3.41}$$

The finite sum $\sum_j |u_j| 1_{|u_j|>1}$ is almost surely finite, so as before it follows that

$$\lim_{\delta\to 0} \limsup_{\epsilon\to 0} \mathbb{P}^\epsilon\big[W_K(\delta)\big(\sum_j |u_j| 1_{|u_j|>1}\big) > \eta\big] = 0. \tag{3.42}$$

By the L_2 isometry of the compensated Poisson integrals,

$$\mathbb{E}\left[\iint_{[-1,1]\times\Omega}\Big(K(x,\omega)-K(y,\omega)\Big)u\tilde{N}^{\epsilon}(du\,d\omega)\right]^2=\iint_{[-1,1]\times\Omega}\Big(K(x,\omega)-K(y,\omega)\Big)^2u^2\;\nu^{\epsilon}(du\,d\omega)$$
$$\leq CW_K^2(|x-y|), \tag{3.43}$$

where $C\equiv\iint_{[-1,1]\times\Omega}u^2\;\nu(du\,d\omega)<\infty$ by (3.31). Let $N(a,\mathcal{X},d)$ denote the covering number of $\mathcal{X}$ with d-balls of radius a, for the semi-metric d on $\mathcal{X}$ defined by

$$d(x,y)\equiv W_K(|x-y|),\;\; x,y\in\mathcal{X}.$$

Since

$$\Big\{y\in\mathcal{X}:d(x,y)\leq\delta\Big\}=\Big\{y\in\mathcal{X}:|x-y|\leq W_k^{\leftarrow}(\delta)\Big\}$$

and $\mathcal{X}\subset\mathbb{R}$ there exists a constant $K_1<\infty$ such that for any $a>0$,

$$N(a,\mathcal{X},d)\leq K_1\frac{1}{W_K^{\leftarrow}(a)}. \tag{3.44}$$

By theorem 2.2.4 of [van der Vaart and Wellner 1996, pg. 98], for a positive constant K independent of ϵ and any $r>0$

$$\mathbb{E}\sup_{|x-y|<\delta}\Big|\iint_{[-1,1]\times\Omega}\Big(K(x,\omega)-K(y,\omega)\Big)u\tilde{N}^{\epsilon}(du\,d\omega)\Big|^2\leq K\Big[\int_0^r\sqrt{N(a,\mathcal{X},d)}\,da+\delta N(r,\mathcal{X},d)\Big]^2 \tag{3.45}$$

Hence under condition B holds true, for small enough r

$$\mathbb{E}\sup_{|x-y|<\delta}\Big|\iint_{[-1,1]\times\Omega}\Big(K(x,\omega)-K(y,\omega)\Big)u\tilde{N}^{\epsilon}(du\,d\omega)\Big|^2\leq K_2\Big[\int_0^r\frac{1}{\sqrt{W_K^{\leftarrow}(a)}}\,da+\delta\frac{1}{W_K^{\leftarrow}(r)}\Big]^2. \tag{3.46}$$

where $K_2<\infty$ is independent of ϵ. The right hand side of the above equation can be made arbitrarily small by first choosing r and then δ. Therefore by Markov's

inequality, for any $\eta > 0$,

$$\lim_{\delta\to 0}\limsup_{\epsilon\to 0}\mathbb{P}\Big(\sup_{|x-y|<\delta}\Big|\iint_{[-1,1]\times\Omega}\Big(K(x,\omega)-K(y,\omega)\Big)u\tilde{N}^{\epsilon}(du\,d\omega)\Big|>\eta\Big) \\ \leq \lim_{\delta,r\to 0}\frac{K_2}{\eta^2}\Big[\int_0^r\frac{1}{\sqrt{W_K^{\leftarrow}(a)}}\,da+\delta\frac{1}{W_K^{\leftarrow}(r)}\Big]^2=0. \tag{3.47}$$

Hence the tightness of the sequence $\{\Pi^{\epsilon}\}$ follows and the theorem is proved. ■

The next theorem is our main result in this section. It gives a simply verifiable sufficient condition for the Lévy random field priors so that posterior consistency holds in most nonparametric regression models.

Recall that $\mathcal{X}$ is a compact subset of $\mathbb{R}$ and Θ_D is

$$\Theta_D \equiv \left\{\theta\in C(\mathcal{X}):\theta(x)\equiv\sum_{j=1}^{J}u_jK(x,\omega_j),\quad J\in\mathbb{N},\ u_j\in\mathbb{R},\ \omega_j\in\Omega.\right\}$$

Let Θ_D^- be its closure in the uniform topology and let $\mathcal{B}(\Theta_D^-)$ be the associated Borel σ-field. For $\theta_0\in\Theta$, denote the "uniform ball" of radius δ centered at θ_0,

$$B_\delta(\theta_0)\equiv\{\theta'\in\Theta_D^-:\|\theta-\theta'\|_*<\delta\} \tag{3.48}$$

where

$$\|\theta\|_*\equiv\sup_{x\in\mathcal{X}}|\theta(x)|. \tag{3.49}$$

Theorem 3.3.3. *Let ν satisfy* (3.31) *and $\mathcal{L}\sim$ Lévy(ν). Let Π be the probability measure on $(\Theta_D^-,(\mathcal{B}(\Theta_D^-)))$ induced by the integral operator $\mathcal{K}$ defined as in* (3.17). *For a uniformly continuous $K(x,\omega):\mathcal{X}\times\Omega\to\mathbb{R}$,*

1. *For any $a>0$,*

$$\iint_{\mathbb{R}\times\Omega}(a\wedge|u|)\ \nu(du\,d\omega)<\infty$$

.

2. Supp $\nu(\cdot) = \mathbb{R} \times \Omega$.

Then $\Pi\big(B_\delta(\theta_0)\big) > 0$ *for every* $\theta_0 \in \Theta_D^-$ *and for all* $\delta > 0$.

Proof. Fix $\delta > 0$ and $\theta_0 \in \Theta_D^-$. Since $\theta_0 \in \Theta_D^-$, there exist a $J_\delta \in \mathbb{N}$ and $\{u_j^*, \omega_j^*\}_{j=1}^{J_\delta}$ such that

$$\|\theta_0(x) - \sum_{j=1}^{J_\delta} K(x, \omega_j^*) u_j^*\|_* < \delta/2. \tag{3.50}$$

Set $\Psi \equiv \sum_{j=1}^{J_\delta} |u_j^*|$, $\kappa \equiv \sup |K(x,\omega)|$ and $\eta \equiv \frac{\delta}{2(\kappa+\Psi)}$. Since K in uniformly continuous in $\mathcal{X} \times \Omega$, there exists an $\eta' > 0$ such that $|\omega - \omega'| < \eta' \Rightarrow |K(x,\omega) - K(x,\omega')| < \eta,\ \forall x \in \mathcal{X}$. Define

$$B'_\delta(\theta_0) \equiv \left\{ \theta : \theta(x) = \sum_{j=1}^{J_\delta} u_j K(x, \omega_j),\ |u_j - u_j^*| < \eta,\ |\omega_j - \omega_j^*| < \eta' \right\}. \tag{3.51}$$

Lemma 3.3.4. $B'_\delta(\theta_0) \subset B_\delta(\theta_0)$

Proof. For any $\theta \in B'_\delta(\theta_0)$ such that $\theta = \sum_{i=1}^{J_\delta} u_j K(x, \omega_j)$,

$$\begin{aligned}
|\theta(x) - \sum_{j=1}^{J_\delta} K(x,\omega_j^*) u_j^*| &\leq \sum_{j=1}^{J_\delta} |K(x,\omega_j) u_j - K(x,\omega_j^*) u_j^*| \\
&\leq \sum_{j=1}^{J_\delta} |K(x,\omega_j) u_j - K(x,\omega_j) u_j^*| + \sum_{j=1}^{J_\delta} |K(x,\omega_j) u_j^* - K(x,\omega_j^*) u_j^*| \\
&\leq \kappa \sum_{j=1}^{J_\delta} |u_j - u_j^*| + \sum_{j=1}^{J_\delta} |u_j^*| \eta \\
&\leq \kappa\eta + \Psi\eta = \delta/2.
\end{aligned}$$

Hence $\|\theta - \sum_{j=1}^{J_\delta} K(x,\omega_j^*) u_j^*\|_* < \delta/2$ and by the triangle inequality $\|\theta - \theta_0\|_* < \delta$. Hence $B'_\delta(\theta_0) \subset B_\delta(\theta_0)$ and the lemma is proved. ■

Pick $a_\delta \leq \inf |u_j^*|$ such that

$$\mathbb{P}\left(\iint_{(-a_\delta,a_\delta)\times\Omega} |K(x,\omega)||u|N(du\,d\omega)| > \delta/2\right) \;\leq\; \frac{2}{\kappa\delta}\iint_{(0,a_\delta)\times\Omega} |u|\nu(du,d\omega < 1. \tag{3.52}$$

$$\begin{aligned}
\Pi^\epsilon\big(B_\delta(\theta_0)\big) &= \mathbb{P}\left(\int_\Omega K(x,\omega)\; uN(du\,d\omega) \in B_\delta^C(\theta_0)\right) \\
&= \mathbb{P}\left(\sum_{j=1}^\infty K(x,\omega_j)u_j \in B_\delta(\theta_0)\right) \geq \mathbb{P}\left(\sum_{j=1}^\infty K(x,\omega_j)u_j \in B'_\delta(\theta_0)\right) \\
&\geq \mathbb{P}\left(\Big\{(u_j,\omega_j)_{j=1}^{J_\delta} : |u_j - u_j^*| < \eta,\; |\omega_j - \omega_j^*| < \eta'\Big\} \bigcap \Big\{\sum_{j=1}^\infty |u_j|1_{|u|_j<a_\delta} < \frac{\delta}{2\kappa}\Big\}\right)
\end{aligned}$$

Using the independent increments property of the Lévy random measures, the quantity on the right hand side of the last equation can be written as

$$= \mathbb{P}\Big(\Big\{\sum_{j=1}^\infty |u_j|1_{|u|_j<a_\delta} < \frac{\delta}{2\kappa}\Big)\frac{\nu(A_\delta)^{J_\delta}e^{-A_\delta}}{J_\delta!}\prod_{j=1}^{J_\delta}\left[\frac{\nu\left\{|u_j - u_j^*| < \eta,\; |\omega_j - \omega_j^*| < \eta'\right\}}{\nu(A_\delta)}\right] \tag{3.53}$$

where $A_\delta \equiv (-a_\delta, a_\delta)\times\Omega$. Since the support of ν is $\mathbb{R}\times\Omega$,

$$\nu\Big\{(u_j,\omega_j) : |u_j - u_j^*| < \eta,\; |\omega_j - \omega_j^*| < \eta'\Big\} > 0. \tag{3.54}$$

Therefore by (3.52) and (3.54) $\Pi\big(B_\delta(\theta_0)\big) > 0$ and the theorem follows. ■

Remark 3.3.5. *Hypothesis (2) of theorem 3.3.3 (i.e., Supp(ν) = $\mathbb{R}\times\mathcal{X}$) can be weakened by the following condition: $\mathbb{R} = G[Supp\ \nu(du\,\omega), +]$ for each $\omega\in\Omega$, where $G[Supp\ \nu(du\,\omega), +]$ is the additive group generated by the set Supp $\nu(du\,\omega)$.*

Remark 3.3.6. *The theorem holds even if the Lévy measure does not satisfy hypothesis (1), but the proof requires a different technique and some extra assumptions on the smoothness of the kernel K are needed.*

Remark 3.3.7. *If Ω is bounded (not necessarily compact) then the uniform continuity of K implies that $\kappa = \sup K(x, \omega) < \infty$.*

Remark 3.3.8. *The family of functions*

$$K(x, \omega, \phi) \equiv \exp\Big\{ -\phi|x - \omega|^2 \Big\},\ \phi \in \mathbb{R}_+,\ \omega \in \mathbb{R}$$

satisfy the conditions A and B if ϕ and ω are bounded above. However, it can be shown that our main theorem holds even if ϕ and ω are not bounded above. See Tokdar and Ghosh [2007] for related discussion.

3.4 Examples

In this section we consider some concrete examples in which the results from the previous section can be applied.

3.4.1 Poisson Regression

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$. For any vector $k = (k_1, k_2, \cdots, k_d) \in \mathbb{N}^d$, let D^k denote the differential operator,

$$D^k \equiv \frac{\partial^{k_+}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}},$$

where $k_+ \equiv \sum_{i=1}^d k_i$. Denote by $\lfloor \alpha \rfloor$ the greatest integer less than or equal to any $\alpha \in \mathbb{R}_+$. For a function $f : \mathcal{X} \mapsto \mathbb{R}$, set

$$\|f\|_\alpha \equiv \max_{k \le \lfloor \alpha \rfloor} \sup_{x \in \mathcal{X}} |D^k f(x)| + \max_{k \le \lfloor \alpha \rfloor} \sup_{x, y \in \mathcal{X}} \frac{|D^k f(x) - D^k f(y)|}{||x - y||^{\alpha - \lfloor \alpha \rfloor}}, \tag{3.55}$$

where the supremum is taken over all x, y in the interior of $\mathcal{X}$ with $x \neq y$. Let $C^\alpha(\mathcal{X})$ be the set of all continuous functions $f : \mathcal{X} \mapsto \mathbb{R}$ with $\|f\|_\alpha < \infty$ and, for $\gamma > 0$, set

$$\Theta_\gamma \equiv \{\theta \in C^\alpha(\mathcal{X}) : \quad (\forall x)\ \theta(x) \ge \gamma\},$$

the set of C^α functions bounded below uniformly by $\gamma > 0$, with Borel σ-field $\mathcal{B}(\Theta_\gamma)$. We consider the following Poisson regression model:

$$\mathbb{P}(Y_i = y|X_i, \theta) = \exp\big(-\theta(X_i)\big)\frac{\theta(X_i)^y}{y!}, \qquad \theta \in \Theta_\gamma,\ y \in \mathbb{Z}_+,\ i \in \mathbb{N}. \tag{3.56}$$

For $\theta \in \Theta_\gamma$ let $\mathbb{P}_\theta$ be the distribution of conditionally-independent $\{Y_i\}$ generated according to the model (4.2) with mean function θ. Let $\|\cdot\|_*$ denote the uniform norm in Θ_γ, and for any $\theta \in \Theta_\gamma$ and $\delta > 0$ define

$$B_\delta(\theta) \equiv \{\tilde{\theta} \in \Theta_\gamma : \|\tilde{\theta} - \theta\|_* < \delta\}.$$

For any increasing sequence of positive numbers $\{M_n\}$ (we will choose a specific sequence in Section 4.2.2 below), set

$$\Theta_\gamma^n \equiv \{\theta \in \Theta_\gamma : \|\theta\|_\alpha \leq M_n\}. \tag{3.57}$$

Lévy Random Field Priors

Let $K(x, \omega) : \mathcal{X} \times \Omega \mapsto \mathbb{R}$ be a Borel measurable function on $\mathcal{X} \times \Omega$ such that the set

$$\left\{\sum_{n=1}^{N} K(x, \omega_j)u_j,\ \{u_j, \omega_j\} \in \mathbb{R} \times \Omega,\ N \in \mathbb{N}\right\}$$

is dense in $C(\mathcal{X})$ in the uniform topology. Let ν be a positive sigma-finite measure on $\mathbb{R} \times \Omega$ satisfying the integrability condition (3.31). Consider the following prior $\Pi_L(d\theta)$ on Θ_γ:

$$\theta(x) = \int_\Omega K(x, \omega)\mathcal{L}(d\omega) \tag{3.58}$$

$$\mathcal{L}(d\omega) \sim \text{Lévy}(\nu). \tag{3.59}$$

Let $\Pi_L(\cdot|Y_1, Y_2, \cdots, Y_n)$ denote the corresponding posterior conditioned on the observations $\{Y_i\}$ sampled according to the Poisson regression model given in (4.2).

Theorem 3.4.1. *If either condition A or B holds true, for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pi_L \left\{ \left[\theta \in \Theta_\gamma^n : \int_{\mathcal{X}} |\theta(x) - \theta_0(x)| \, Q_n(dx) > \epsilon \right] \Big| Y_1, Y_2, \cdots, Y_n \right\} = 0 \qquad (3.60)$$

in $\mathbb{P}_{\theta_o}$ probability.

Proof. If condition A or B holds true, by theorem (3.3.3) it follows that $\Pi_L\big(B_\delta(\theta_0)\big) > 0$ for any $\theta_0 \in \Theta_\gamma$ and $\delta > 0$. Hence the result follows from theorem (4.2.1). ■

3.4.2 Regression with Gaussian Errors

Let $\Theta \equiv C^\alpha(\mathcal{X}), \alpha > 0$. We consider the following regression model on Θ. Let $X_i \in \mathcal{X} \subset \mathbb{R}$ and

$$Y_i \big| \theta, X_i = \theta(X_i) + \epsilon_i, \ \theta_0 \in \Theta, \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2). \qquad (3.61)$$

Here the error variance $\sigma > 0$ is assumed to be unknown. Let $\Pi_\sigma(d\sigma)$ denote a (prior) probability distribution whose support is the postive real line. Let $\Pi_L(d\theta)$ be the Lévy random field prior on Θ defined in (3.58). Let $\Pi_L(d\theta) \otimes \Pi_\sigma(d\sigma)$ be the prior distribution on $\Theta \times \mathbb{R}_+$ and $\Pi_{L,\sigma}(\cdot | Y_1, Y_2, \cdots, Y_n)$ denote the corresponding posterior distribution. Define the sets

$$\Theta^n \equiv \big\{ \theta \in \Theta : \|\theta\|_\alpha \le n^\beta, \ \beta \in (1/2, 1] \big\}$$

and set $S_{n,\epsilon}$

$$S_{n,\epsilon} \equiv \Big\{ \theta \in \Theta^n : \int_{\mathcal{X}} |\theta(x) - \theta_0(x)| \, Q_n(dx) < \epsilon, \ \Big| \frac{\sigma}{\sigma_0} - 1 \Big| < \epsilon \Big\}, \ Q_n(dx) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(dx). \qquad (3.62)$$

Theorem 3.4.2. *Let $\mathbb{P}_{\theta_0,\sigma_0}$ denote the distribution of $\{Y_i\}$ conditional on (θ_0, σ_0). If conditions A or B hold true, for any $\theta_0 \in \Theta$ and for any $\epsilon > 0$*

$$\lim_{n\to\infty} \Pi_{L,\sigma}\Big(S_{n,\epsilon}^C \ \Big| \ Y_1, Y_2, \cdots, Y_n \Big) = 0$$

in $\mathbb{P}_{\theta_0,\sigma_0}$ probability.

Proof. Direct calculations yield that (also see [Choi and Schervish 2007, pg. 1974])

$$K_i(\theta_0,\theta) = \frac{1}{2}\log\frac{\sigma^2}{\sigma_0^2} - \frac{1}{2}\left(1-\frac{\sigma_0^2}{\sigma^2}\right) + \frac{1}{2}\frac{\left[\theta_0(X_i)-\theta(X_i)\right]^2}{\sigma^2}$$

and

$$V_i(\theta_0,\theta) = \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma^2}-1\right)^2 + \frac{\sigma_0^4}{\sigma^4}\Big(\theta_0(X_i)-\theta(X_i)\Big)^2.$$

Define

$$B_\delta^{\theta_0,\sigma_0} \equiv \left\{(\theta,\sigma) : \|\theta-\theta_0\|_* < \delta, \left|\frac{\sigma}{\sigma_0}-1\right| < \delta\right\}.$$

Hence the above calculations show that for every $\epsilon > 0$, there exists a $\delta > 0$ such that for all $(\theta,\sigma) \in B_\delta^{\theta_0,\sigma_0}$,

$$K_i(\theta_0,\theta) < \epsilon \ \forall i, \ \sum_{i=1}^{\infty}\frac{V_i(\theta_0,\theta)}{i^2} < \infty.$$

Hence conditions I.1 and I.2 are satisfied if the prior probability of the set $B_\delta^{\theta_0,\sigma_0}$ is positive for every $\delta > 0$. If either condition A or B is satisfied, theorem (3.3.3) gaurantees that $\Pi_L\big(B_\delta(\theta_0)\big) > 0$. Since the prior $\Pi_\sigma(d\sigma)$ has support over the entire real line, $\Pi_\sigma\Big(\Big|\frac{\sigma}{\sigma_0}-1\Big| < \delta\Big) > 0$. Hence for every $\delta > 0$, $\Pi_{L,\sigma}\Big(B_\delta^{\theta_0,\sigma_0}\Big) > 0$ and conditions I.1 and I.2 are satisfied.

The existence of tests as required in II.1 and II.2 are constructed in [Choi and Schervish 2007, Theorem 2, pg. 1973]. Hence the theorem is proved. ■

Remark 3.4.3. *If the covariates X_i were sampled uniformly from $\mathcal{X}$ and σ_0 is known, it follows that*

$$K_i(\theta_0,\theta) = \frac{1}{2\sigma_0^2}\ \|\theta_0-\theta\|_{L_2(\mathcal{X})} \le \frac{1}{2\sigma_0^2}\ \|\theta_0-\theta\|_* \tag{3.63}$$

Hence the results of the previous case hold in this case as well. However, condition B for the function $K(\cdot,\cdot)$ is too strong. Notice that theorem (3.3.3) *gives that* $\Pi_L\big(\theta : \|\theta_0 - \theta\|_* < \delta\big) > 0$ *whereas we only need* $\Pi_L\big(\theta : \|\theta_0 - \theta\|_{L_2(\mathcal{X})} < \delta\big) > 0$. *To show this and thereby relax the assumptions on condition B, we will have to show the weak convergence of probability measures on* $L_2(\mathcal{X})$ *(instead of* $C(\mathcal{X})$ *as in theorem* (3.3.2)*) induced by the truncated compound Poisson Lévy measures.*

Chapter 4

Posterior Consistency of Nonparametric Poisson Regression Models

4.1 Introduction

Posterior consistency for non-identically distributed random variables has received attention only recently in the literature. The general framework for proving such results was laid out by Amewou-Atisso *et al.* [2003], Ghosal and van der Vaart [2007a,b] and Choi and Schervish [2007]. Choi and Schervish [2007] prove the posterior consistency for a regression model with Gaussian error terms. Recently posterior consistency was proved for nonparametric binary regression by Ghosal and Roy [2006], who conjectured that similar results might hold for the Poisson regression.

In this note we verify that conjecture, proving posterior consistency for the Poisson regression model under moderate assumptions. The standard proof technique requires one to verify two conditions: positivity of prior mass for all information-metric (or equivalently Kullback-Leibler) neighborhoods of the true parameter, and existence of point-separating hypothesis tests with exponentially decaying error rates of types I and II. The first condition is prior-specific, and is relatively easy to verify. The existence of tests is independent of the prior and usually requires the calculation of

the metric entropy of the parameter space. Our main tool in verifying the tests is a condition for exponential convergence rates for the law of large numbers, derived in Baum *et al.* [1962].

4.2 Main Result

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$. For any vector $k = (k_1, k_2, \cdots, k_d) \in \mathbb{N}^d$, let D^k denote the differential operator,

$$D^k \equiv \frac{\partial^{k_+}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}},$$

where $k_+ \equiv \sum_{i=1}^d k_i$. Denote by $\lfloor \alpha \rfloor$ the greatest integer less than or equal to any $\alpha \in \mathbb{R}_+$. For a function $f : \mathcal{X} \mapsto \mathbb{R}$, set

$$\|f\|_\alpha \equiv \max_{k \leq \lfloor \alpha \rfloor} \sup_{x \in \mathcal{X}} |D^k f(x)| + \max_{k \leq \lfloor \alpha \rfloor} \sup_{x,y \in \mathcal{X}} \frac{|D^k f(x) - D^k f(y)|}{||x - y||^{\alpha - \lfloor \alpha \rfloor}}, \tag{4.1}$$

where the supremum is taken over all x, y in the interior of $\mathcal{X}$ with $x \neq y$. Let $C^\alpha(\mathcal{X})$ be the set of all continuous functions $f : \mathcal{X} \mapsto \mathbb{R}$ with $\|f\|_\alpha < \infty$ and, for $\gamma > 0$, set

$$\Theta_\gamma \equiv \{\theta \in C^\alpha(\mathcal{X}) : \quad (\forall x)\ \theta(x) \geq \gamma\},$$

the set of C^α functions bounded below uniformly by $\gamma > 0$, with Borel σ-field $\mathcal{B}(\Theta_\gamma)$. We consider the following Poisson regression model:

$$\mathbb{P}(Y_i = y | X_i, \theta) = \exp\big(-\theta(X_i)\big) \frac{\theta(X_i)^y}{y!}, \qquad \theta \in \Theta_\gamma,\ y \in \mathbb{Z}_+,\ i \in \mathbb{N}. \tag{4.2}$$

For $\theta \in \Theta_\gamma$ let $\mathbb{P}_\theta$ be the distribution of conditionally-independent $\{Y_i\}$ generated according to the model (4.2) with mean function θ. Let $\|\cdot\|_*$ denote the uniform norm in Θ_γ, and for any $\theta \in \Theta_\gamma$ and $\delta > 0$ define

$$B_\delta(\theta) \equiv \{\tilde{\theta} \in \Theta_\gamma : \|\tilde{\theta} - \theta\|_* < \delta\}.$$

For any increasing sequence of positive numbers $\{M_n\}$ (we will choose a specific sequence in Section 4.2.2 below), set

$$\Theta_\gamma^n \equiv \{\theta \in \Theta_\gamma : \|\theta\|_\alpha \leq M_n\}. \tag{4.3}$$

Theorem 4.2.1. *Let Π be a (prior) probability measure on $\mathcal{B}(\Theta_\gamma)$ such that $\Pi\big(B_\delta(\theta)\big) > 0$ for each $\theta \in \Theta_\gamma$ and $\delta > 0$. Then for any $\epsilon > 0$, and any $\theta_0 \in \Theta_\gamma$,*

$$\lim_{n\to\infty} \Pi\left(\left\{\theta \in \Theta_\gamma^n : \int_{\mathcal{X}} |\theta(x) - \theta_0(x)|\, Q_n(dx) > \epsilon\right\} \middle| Y_1, Y_2, \cdots, Y_n\right) = 0$$

in $\mathbb{P}_{\theta_0}$ probability, where $Q_n(dx) \equiv \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(dx)$ denotes the empirical distribution.

Proof. Fix $\epsilon > 0$ and $\theta_0 \in \Theta_\gamma$, and let

$$\begin{aligned}\Lambda_i(\theta_0, \theta) &\equiv \log \frac{\exp\big(-\theta_0(X_i)\big)\big(\theta_0(X_i)\big)^{Y_i}}{\exp\big(-\theta(X_i)\big)\big(\theta(X_i)\big)^{Y_i}} \\ &= \big[\theta(X_i) - \theta_0(X_i)\big] + Y_i \log \frac{\theta_0(X_i)}{\theta(X_i)}\end{aligned}$$

denote the log ratio of two Poisson densities with means $\theta_0(X_i)$ and $\theta(X_i)$, evaluated at Y_i. Set

$$K_i(\theta_0, \theta) \equiv \mathbb{E}_{\theta_0}\big(\Lambda_i(\theta_0, \theta)\big)$$

(the Kullback-Leibler divergence for observation i), and

$$V_i(\theta_0, \theta) \equiv \mathsf{Var}_{\theta_0}\big(\Lambda_i(\theta_0, \theta)\big).$$

Also set

$$U_n^\epsilon \equiv \left\{\theta \in \Theta_\gamma : \int |\theta(x) - \theta_0(x)|\, Q_n(dx) > \epsilon\right\}.$$

Our goal is to verify the following two conditions:

I) Positive prior probability: There exists a set $B \subset \Theta_\gamma$ with $\Pi(B) > 0$ such that

1. $(\forall \theta \in B)\ \sum_{i=1}^{\infty} V_i(\theta_0, \theta)/i^2 < \infty$.

2. $(\forall i \in \mathbb{N})\ \Pi\big(B \cap \{\theta : K_i(\theta_0, \theta) < \epsilon\}\big) > 0$.

II) Existence of tests: There exist non-negative bounded measurable functions $\{\Phi_n\}_{n=1}^{\infty}$ and positive constants $C_1, c_1 > 0$ such that

1. $\sum_{n=1}^{\infty} \mathbb{E}_{\theta_0} \Phi_n < \infty$.

2. $(\forall \theta \in U_n^{\epsilon} \cap \Theta_{\gamma}^n)\ \mathbb{E}_{\theta}(1 - \Phi_n) \leq C_1 e^{-c_1 n}$.

By Theorem 1 of [Choi and Schervish 2007], these conditions imply that

$$\lim_{n \to \infty} \Pi\big(U_n^{\epsilon} \cap \Theta_{\gamma}^n | Y_1, Y_2, \cdots, Y_n\big) = 0, \quad \text{in } \mathbb{P}_{\theta_0} \text{ probability.}$$

4.2.1 Positive prior probability

First we consider Condition I. Fix any $\delta_0 > 0$ and set $B \equiv B_{\delta_0}(\theta_0)$. Then for any $\theta \in B$,

$$V_i(\theta_0, \theta) = \theta_0(X_i) \left[\log \frac{\theta_0(X_i)}{\theta(X_i)}\right]^2 \leq \|\theta_0\|_* \left[\log \frac{\|\theta_0\|_* + \delta_0}{\gamma}\right]^2,$$

uniformly in $i \in \mathbb{N}$, so $\sum_{i=1}^{\infty} V_i(\theta_0, \theta)/i^2 < \infty$, for all $\theta \in B$ and I.1 holds.

Notice that

$$\begin{aligned} K_i(\theta_0, \theta) &= \theta(X_i) - \theta_0(X_i) + \theta_0(X_i) \log \frac{\theta_0(X_i)}{\theta(X_i)} \\ &= \theta(X_i) \left(1 - r + r \log r\right), \end{aligned}$$

where $r \equiv \frac{\theta_0(X_i)}{\theta(X_i)}$. The inequality $\log r \leq r - 1$ implies that

$$\begin{aligned} K_i(\theta_0, \theta) &\leq \theta(X_i)\big(1 - r + r(r-1)\big) = \theta(X_i)(r-1)^2 \\ &= \frac{\big(\theta(X_i) - \theta_0(X_i)\big)^2}{\theta(X_i)} \leq \frac{1}{\gamma} \|\theta - \theta_0\|_*^2. \end{aligned}$$

Since $\Pi\big(B_{\delta}(\theta_0)\big) > 0$ for each $\delta > 0$ by hypothesis, and in particular for $\delta = \big(\delta_0 \wedge \sqrt{\gamma \epsilon}\big)$, condition I.2 is verified since $B_{\delta}(\theta_0) \subset B \cap \{\theta : K_i(\theta_0, \theta) < \epsilon\}$.

4.2.2 Existence of tests

Now we turn to the second condition. First we construct tests for the simple *vs.* simple hypotheses:

$$H_0 : Y_i \sim \mathbf{Po}\big(\theta_0(X_i)\big) \text{ vs. } H_1 : Y_i \sim \mathbf{Po}\big(\theta_1(X_i)\big).$$

Below we construct test functions for testing a more general class of hypotheses. Before proceeding, we state a result of Baum *et al.* [1962] on the exponential convergence rates for the law of large numbers used later in our proof.

Theorem 4.2.2. *[Baum* et al. *1962, Theorem 2] Let* $\{X_k, k = 1, 2, \cdots\}$ *be a sequence of independent random variables and let* $S_n = \sum_{k=1}^{n} X_k$*. For every* $\epsilon > 0$*, and a constant* c *there exists a* $\rho_\epsilon \in (0, 1)$ *such that*

$$\mathbb{P}\Big(\Big|\frac{S_n}{n} - c\Big| \geq \epsilon\Big) \leq C\rho_\epsilon^n \tag{4.4}$$

if and only if there exists a constant M_ϵ *and* $t_\epsilon > 0$ *such that*

$$\mathbb{E}\Big(e^{t(S_n - nc)}\Big) \leq M_\epsilon\, e^{|t|\, n\, \epsilon}, \quad \forall t \in [-t_\epsilon, t_\epsilon]. \tag{4.5}$$

Lemma 4.2.3. *Let* $\{Y_i\}$ *be independent Poisson random variables with means* μ_i *and let* $\{X_i\} \subset \mathcal{X}$*. Consider the following hypothesis testing problem:*

$$H_0 : \mu_i = \theta_0(X_i) \text{ vs. } H_1 : \mu_i = \theta_1(X_i)$$

for fixed elements $\theta_0, \theta_1 \in \Theta_\gamma$ *that satisfy the inequality*

$$\theta_1(x) > \theta_0(x) + \epsilon$$

for some number $\epsilon > 0$ *and all* $x \in \mathcal{X}$*. Consider the sequence of indicator random variables,*

$$\Phi_n \equiv \mathbf{1}_{\{A_n\}}, \qquad A_n \equiv \left\{\frac{1}{n}\sum_{i=1}^{n}\big[Y_i - \theta_0(X_i)\big] > \epsilon/2\right\}.$$

Then, for every $\epsilon > 0$ and $n \in \mathbb{N}$, there exists a $\rho_\epsilon \in (0,1)$ such that

$$\mathbb{E}_{\theta_0}(\Phi_n) \leq \exp(n \ \log \rho_\epsilon) \tag{4.6a}$$

$$\mathbb{E}_{\theta_1}(1 - \Phi_n) \leq \exp(n \ \log \rho_\epsilon). \tag{4.6b}$$

Proof. The proof is deferred to the Appendix. ■

Remark 4.2.4. *In the above lemma if we had, $\theta_0(x) > \theta_1(x) + \epsilon$, with all the other conditions being the same, then the test*

$$\Phi_n = \mathbf{1}_{\{B_n\}}, \ B_n \equiv \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \theta_1(X_i) \right] < \epsilon/2 \right\}$$

will satisfy (4.6), for every $\epsilon > 0$ and $n \in \mathbb{N}$. (see Appendix for the details of the proof).

The rest of our proof is based on a standard argument using covering numbers [Ghosal and Roy 2006]. First notice that for any $\theta_j \in \Theta_\gamma$ with $\int |\theta_j(x) - \theta_0(x)| Q_n(dx) > \epsilon$,

$$Q_n\{x : |\theta_j(x) - \theta_0(x)| > \epsilon/2\} \geq \frac{\epsilon}{2(\|\theta_0\|_* + \|\theta_j\|_*)}.$$

This can be seen from:

$$\begin{aligned} \epsilon &< \int |\theta_j(x) - \theta_0(x)| \, Q_n(dx) \\ &< \left(\|\theta_0\|_* + \|\theta_j\|_* \right) Q_n\{x : |\theta_j(x) - \theta_0(x)| > \epsilon/2\} + \epsilon/2. \end{aligned}$$

Hence the cardinality $\#\{i : |\theta_j(X_i) - \theta_0(X_i)| > \epsilon/2\} \geq \tilde{C}n$ for the constant $\tilde{C} \equiv \epsilon(2\|\theta_0\|_* + 2\|\theta_j\|_*)^{-1}$. This implies that one of the two sets $\{i : \theta_j(X_i) > \theta_0(X_i) + \epsilon/2\}$

or $\{i : \theta_0(X_i) > \theta_j(X_i) + \epsilon/2\}$ will have at least $\tilde{C}n/2$ points. For definiteness we will assume it is the former and set

$$S^+_{\theta_j} \equiv \{x : \theta_j(x) > \theta_0(x) + \epsilon/2\} \tag{4.7}$$

(otherwise see Remark 4.2.4, and make the corresponding change in equation (4.8) below). Now for testing θ_0 *vs.* θ_j such that $\int |\theta_j(x) - \theta_0(x)| \, Q_n(dx) > \epsilon$, we construct the test:

$$\Psi_{nj} = \mathbf{1}_{\{E_{nj}\}}, \;\; E_{nj} \equiv \left\{ \frac{1}{n} \sum_{i=1}^{n} \big[Y_i - \theta_0(X_i)\big] \mathbf{1}_{\{X_i \in S^+_{\theta_j}\}} > \epsilon/2 \right\}. \tag{4.8}$$

Let $N \equiv N\big(\epsilon/2, \Theta^n_\gamma, \|\cdot\|_*\big)$ denote the covering number of the set Θ^n_γ under the uniform norm, *i.e.*, the smallest number N such that

$$\Theta^n_\gamma \subset \bigcup_{j=1}^{N} B_{\epsilon/2}(\theta_j)$$

for some $\{\theta_1, \theta_2, \cdots, \theta_N\} \subset \Theta^n_\gamma$. Then by Theorem 2.7.1 of van der Vaart and Wellner [1996],

$$\log N \leq K_1 M_n^{d/\alpha} \epsilon^{-1/\alpha} \tag{4.9}$$

for some constant $K_1 < \infty$. For each $\theta \in \Theta^n_\gamma$ there is some $j \leq N$ such that $\|\theta - \theta_j\|_* \leq \epsilon/2$. Recall the tests $\{\Psi_{nj}\}_{j=1}^N$, from (4.8). By Lemma 4.2.3 (with $\epsilon/2$ replacing ϵ),

$$\begin{aligned} \mathbb{E}_{\theta_0}(\Psi_{nj}) &\leq \exp(n \log \rho_\epsilon), \\ \mathbb{E}_{\theta_j}(1 - \Psi_{nj}) &\leq \exp(n \log \rho_\epsilon), \end{aligned} \qquad 1 \leq j \leq N.$$

Now set $\Psi_n \equiv \max\{\Psi_{n,\theta_j} : \; 1 \leq j \leq N\}$. Then,

$$\mathbb{E}_{\theta_0}(\Psi_n) \leq \sum_{j=1}^{N} \mathbb{E}_{\theta_j}(\Psi_{nj}) \leq \exp(\log N + n \log \rho_\epsilon). \tag{4.10}$$

If we set $M_n \equiv K_2 n^{\alpha/d}$ (see equation (4.3)) for sufficiently small K_2, then equations (4.9, 4.10) imply

$$\mathbb{E}_{\theta_0}(\Psi_n) \leq \exp(-K_3 n)$$

for some constant $K_3 > 0$. In particular, $\sum_{n=1}^{\infty} \mathbb{E}_{\theta_0}(\Psi_n) < \infty$, verifying II.1. Now we turn to II.2.

For any $\theta \in U_n^\epsilon \cap \Theta_\gamma^n$, there is at least one $j \in \{1, 2, \cdots, N\}$ such that $\|\theta - \theta_j\|_* < \epsilon/2$. For $X_i \in S_{\theta_j}^+$ (see equation (4.7)), by the triangle inequality, $\left[\theta(X_i) - \theta_0(X_i)\right] \geq \epsilon/2$. Hence

$$\mathbb{E}_\theta(1 - \Psi_n) \leq \mathbb{E}_\theta(1 - \Psi_{nj}) \leq \exp(n \log \rho_\epsilon).$$

Since this holds uniformly for all $\theta \in U_n^\epsilon \cap \Theta_\gamma^n$,

$$\sup_{\theta \in U_n^\epsilon \cap \Theta_\gamma^n} \mathbb{E}_\theta(1 - \Psi_n) \leq \exp(n \log \rho_\epsilon)$$

verifying II.2. Hence we have proved the theorem. ■

Remark 4.2.5. *If the prior* Π *also satisfies the condition* $\Pi(\Theta_\gamma^{nc}) \leq C_2 e^{-c_2 n}$ *for constants* $C_2, c_2 > 0$*, then the conclusion of Theorem 4.2.1 can be strengthened to:*

$$\lim_{n \to \infty} \Pi \left\{ \left[\theta \in \Theta_\gamma : \int_{\mathcal{X}} |\theta(x) - \theta_0(x)| \, dQ_n(x) > \epsilon \right] \Big| Y_1, Y_2, \cdots, Y_n \right\} = 0$$

in $\mathbb{P}_{\theta_0}$ *probability.*

Remark 4.2.6. *The predictors* $\{X_i\}$*, instead of being fixed, can be sampled according to a distribution* D *on* $\mathcal{X}$*. With similar arguments as above, we can show that consistency holds in* L_1 *norm, instead of the empirical* L_1 *norm. See [Ghosal and Roy 2006] for a result of this type.*

4.3 An Example

Set $\mathcal{X} \equiv [-1, 1]$. Fix $\alpha, \gamma > 0$ and set $\Theta_\gamma \equiv \{\theta \in C^\alpha(\mathcal{X}) : \theta(x) \geq \gamma\}$. Let $\Phi \equiv \{\phi_k(x),\ k \in \mathbb{N},\ x \in \mathcal{X}\}$ denote a set of continuous functions such that Θ_γ is contained in the linear span of Φ in the uniform norm (*e.g.* $\phi_k(x) \equiv P_k(x) = \frac{1}{2^k k!}\frac{d^k}{dx^k}\left[(x^2-1)^k\right]$, the k^{th} Legendre polynomial). Fix $\lambda > 0$, and consider the following prior $\Pi(d\theta)$ on Θ_γ:

$$J \sim \mathrm{Po}(\lambda).$$

$$\{\beta_j\}_{j=1}^J \overset{iid}{\sim} \mathbf{No}(0,1),\ \{\beta_j\}_{j=J+1}^\infty = 0.$$

$$\theta(x) = \sum_{j=1}^\infty \beta_j \phi_j(x).$$

Lemma 4.3.1. *The prior $\Pi(d\theta)$ defined above satisfies $\Pi\big(B_\delta(\theta_0)\big) > 0$ for any $\delta > 0$ and $\theta_0 \in \Theta_\gamma$. Hence the conclusion of Theorem 4.2.1 holds true.*

Proof. Fix $\delta > 0$. For any $\theta_0 \in \Theta_\gamma$, by hypothesis, there exist a $J_\delta \in \mathbb{N}$ and $\{\beta_j^*\}_{j=1}^{J_\delta} \in \mathbb{R}$ such that

$$\|\theta_0(\mathcal{X}) - \sum_{j=1}^{J_\delta} \beta_j \phi_j(x)\|_* \leq \delta/2. \tag{4.11}$$

Let $\kappa \equiv \max_{j \leq J_\delta} \|\phi_j\|_*$. Hence by the triangle inequality and (4.11),

$$B_\delta(\theta_0) \supset \left\{ \sum_{j=1}^\infty \beta_j \phi_j(x) : \sum_{j=1}^{J_\delta} |\beta_j - \beta_j^*| \leq \delta/2\kappa, \{\beta_j\}_{J_\delta+1}^\infty = 0 \right\}. \tag{4.12}$$

The prior probability of the right hand side above is given by

$$\mathbb{P}\left\{ \{\beta_j\}_{j=1}^\infty : \sum_{j=1}^{J_\delta} |\beta_j - \beta_j^*| \leq \delta/2\kappa, \{\beta_j\}_{J_\delta+1}^\infty = 0 \right\} = e^{-\lambda} \frac{1}{(2\pi)^{J_\delta/2}} \int_B \exp(-\beta'\beta/2) d\beta > 0 \tag{4.13}$$

where $B \equiv \left\{ \{\beta_j\}_{j=1}^{J_\delta} : \sum_{j=1}^{J_\delta} |\beta_j - \beta_j^*| \leq \delta/2\kappa \right\}$. Hence $\Pi\big(B_\delta(\theta_0)\big) > 0$ and the lemma is proved. ■

This example shows an illustrates an important fact. While constructing priors based on series expansion, for consistency to hold in Poisson regression models, it is sufficient to verify the following two conditions:

1. The true parameter is contained in the linear span of the basis functions in an appropriate topology.
2. The basis coefficients are supported over the entire real line.

4.4 Discussion

We have proved the posterior consistency of the nonparametric Poisson regression model, under very modest assumptions on the prior: it is sufficient that the prior puts positive mass on the uniform balls around the true parameter. This condition, as demonstrated in the example, is easy to verify for most priors, especially the ones which are based on series expansion or splines. We constructed tests which have the required exponential error rates, and our tests applicable to other exponential error models.

It is interesting to note that we require the parameter space Θ_γ to be bounded away from zero. This is not a mere artifact of our proof: if almost all observations in a sample from a Poisson distribution are $Y_i = 0$, then it hard to estimate the log mean if it is not bounded away from zero. We expect that our approach may be extended to obtain bounds on convergence rates for nonparametric Poisson regression models.

Proof of Lemma 4.2.3. Notice that

$$\mathbb{E}_{\theta_0}(\Phi_n) = \mathbb{P}_{\theta_0}\left\{ \frac{1}{n} \sum_{i=1}^{n} \big[(Y_i - \theta_0(X_i) \big] > \epsilon/2 \right\}.$$

Since under H_0 $\mathbb{E}_{\theta_0}(Y_i) = \theta(X_i)$, by the strong law of large numbers

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \big[(Y_i - \theta_0(X_i)\big] = 0,$$

almost surely. Since under the null hypthesis, the sum $\sum_{i=1}^{n} Y_i$ is a Poisson random variable with mean $\sum_{i=1}^{n} \theta_0(X_i)$,

$$\mathbb{E}\Big[e^{t\left(\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \theta(X_i)\right)}\Big] = \exp\Big((e^t - t - 1) \sum_{i=1}^{n} \theta_0(X_i)\Big) \leq \exp\Big((e^t - t - 1) n \|\theta_0\|_*\Big) \leq \exp\Big(n\epsilon|t|\Big)$$

for all $t \in [-t_\epsilon, t_\epsilon]$, $t_\epsilon \equiv \frac{2\epsilon}{\|\theta_0\|_*}$. Therefore by theorem 4.2.2, there exists a $\rho_\epsilon \in (0,1)$ such that

$$\mathbb{E}_{\theta_0}(\Phi_n) = \mathbb{P}_{\theta_0}\left\{\frac{1}{n}\sum_{i=1}^{n}\big[(Y_i - \theta_0(X_i)\big] > \epsilon/2\right\} \leq (\rho_\epsilon)^n = \exp(n \log \rho_\epsilon)$$

and hence the type I error decays exponentially with n.

To obtain the type II error rate of equation (4.6b), first note

$$\begin{aligned}
\mathbb{E}_{\theta_1}(1 - \Phi_n) &= \mathbb{P}_{\theta_1}\left\{\frac{1}{n}\sum_{i=1}^{n}\big[(Y_i - \theta_0(X_i)\big] \leq \epsilon/2\right\} \\
&= \mathbb{P}_{\theta_1}\left\{\frac{1}{n}\sum_{i=1}^{n}\big[Y_i - \theta_1(X_i)\big] + \frac{1}{n}\sum_{i=1}^{n}\big[\theta_1(X_i) - \theta_0(X_i)\big] \leq \epsilon/2\right\} \\
&\leq \mathbb{P}_{\theta_1}\left\{\frac{1}{n}\sum_{i=1}^{n}\big[Y_i - \theta_1(X_i)\big] \leq -\epsilon/2\right\} \qquad (4.14)
\end{aligned}$$

where the last inequality holds because $\theta_1(X_i) - \theta_0(X_i) > \epsilon$ for each i by hypothesis. Therefore as done in the previous case, theorem 4.2.2 can be applied to show that there exists a $\rho_\epsilon \in (0,1)$ such that

$$\mathbb{P}_{\theta_1}\left\{\frac{1}{n}\sum_{i=1}^{n}\big[Y_i - \theta_1(X_i)\big] \leq -\epsilon/2\right\} \leq \exp(n \log \rho_\epsilon) \qquad (4.15)$$

Hence by (4.14) and (4.15)

$$\mathbb{E}_{\theta_1}(1-\Phi_n) \leq \exp(n\,\log\rho_\epsilon)$$

and therefore the type II error decays exponentially with n and the lemma is proved.

Proof of Remark 4.2.4. In this case the type II error rate follows from theorem 4.2.2,

$$\begin{aligned}\mathbb{E}_{\theta_1}(1-\Phi_n) &= \mathbb{P}_{\theta_1}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\theta_1(X_i)\right] > \epsilon/2\right\}\\ &\leq \exp(n\,\log\rho_\epsilon).\end{aligned}$$

For the type I error,

$$\begin{aligned}\mathbb{E}_{\theta_0}(\Phi_n) &= \mathbb{P}_{\theta_0}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\theta_1(X_i)\right] \leq \epsilon/2\right\}\\ &= \mathbb{P}_{\theta_0}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\theta_0(X_i)\right]+\frac{1}{n}\sum_{i=1}^{n}\left[\theta_0(X_i)-\theta_1(X_i)\right] \geq \epsilon/2\right\}. \qquad (4.16)\end{aligned}$$

Once again the differences $\left[\theta_0(X_i)-\theta_1(X_i)\right] > \epsilon$ for each i, so equation (4.16) can be bounded above by

$$\mathbb{E}_{\theta_0}(\Phi_n) \leq \mathbb{P}_{\theta_0}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\theta_0(X_i)\right] \leq -\epsilon/2\right\} \leq \exp(-n\,\log\rho_\epsilon)$$

by another application of theorem 4.2.2. ■

Chapter 5

Nonparametric Covrariance Function Estimation

5.1 Introduction

In this chapter we discuss a novel application of Lévy random measures for estimating spectral measures and covariance functions of stationary stochastic processes.

Nonparametric Bayesian approaches for estimating a spectral density of a Gaussian time series were studied in [Pawitan and O'Sullivan 1994, Carter and Kohn 1997, Gangopadhyay *et al.* 1999, Choudhuri *et al.* 2004](see Choudhuri *et al.* [2004] for more references). In all of the above papers, the estimation proceeds through first smoothing the periodogram and then using an approximate likelihood (Whittle likelihood) instead of using the actual likelihood for the data which involves the autocorrelations functions. The prior distribution for the spectral density is modeled using a variety of tools such as Brownian motions (Carter and Kohn [1997]), Bernstein polynomials (Choudhuri *et al.* [2004]), piecewise polynomial expansion of the logarithm of the spectral density (Gangopadhyay *et al.* [1999]) *etc.*

Bayesian nonparametric estimation of spatial random effects using a Dirichlet process mixture was studied in Reich and Fuentes [2007]. In Tu *et al.* [2006] the

authors develop Lévy Adaptive Regression Kernels for estimating the mean surface of a spatial random field. In both cases the covariance function is modeled using a standard parametric form.

We focus on nonparametric estimation of the covariance function of a stationary process, and construct a nonparametric prior distribution on the set of positive functions on $\mathbb{R}^d$. Constructing a prior distribution for positive definite functions is challenging (since the positive definiteness of an arbitrary function is difficult to verify) and efficient sampling methods need to be developed for implementation. In this chapter, we illustrate that Bochner's theorem together with Lévy random measures can be used for constructing efficient, easy to implement prior distributions for spectral measures and covriance functions. We consider a few examples mainly in the spatial setting, however our discussion equally applies to time series data as well.

5.2 Spectral representation and priors on stationary covariance functions

Let $S \subset \mathbb{R}^d$ be a spatial domain and let $\{Z(s) : s \in S\}$ be a real valued stationary process. Let $C(h), h \in \mathbb{R}^d$ denote the covariance function of the process $Z(\cdot)$:

$$C(s_2 - s_1) \equiv \mathrm{Cov}\big(Z(s_1), Z(s_2)\big),\ s_1, s_2 \in \mathbb{R}^d. \tag{5.1}$$

To be a valid covariance function, the function $C(h)$ needs to be positive semi-definite, *i.e.,* for any set of locations $\{s_i\}_{i=1}^n$ and constants $\{a_i\}_{i=1}^n$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(s_i - s_j) \geq 0, \quad n \in \mathbb{N},\ s_i, s_j \in S,\ a_i, a_j \in \mathbb{R}. \tag{5.2}$$

Our goal is to define a probability distribution on the set of real valued positive definite functions on S. It is often difficult to verify the postive definiteness of a

function. Bochner's theorem [Bochner 1955, Cramér 1940] gives a necessary and sufficient condition for a function to be positive definite.

Definition 5.2.1. *For a finite Borel measure μ on $\mathbb{R}^d$, define its Fourier transform to be:*

$$\hat{\mu}(\omega) \equiv \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp\big(-ix\cdot\omega\big)\mu(dx), \quad \omega \in \mathbb{R}^d \tag{5.3}$$

where $x \cdot \omega$ denotes the usual Euclidean inner product in $\mathbb{R}^d$.

Theorem 5.2.2. *[Bochner 1955] A real valued function $f(\cdot)$ on $\mathbb{R}^d$ is of positive semi-definite if and only if it is the Fourier transform of a finite, positive, symmetric measure μ on $\mathbb{R}^d$.*

The measure μ above is referred to be the spectral measure. If $\mu(d\omega)$ has a density $f(\omega)$ with respect to the Lebesgue measure, $f(\omega)$ is called the spectral density.

Let $M(\mathbb{R}^d)$ denote the set of positive finite measures on $\mathbb{R}^d$ and $\mathcal{M}(\mathbb{R}^d)$ denote the Borel sigma field under the weak star topology. Hence by Bochner's theorem,

$$\begin{aligned} C(h) &\equiv \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp\big(-ih\cdot\omega\big)\mu(d\omega); \quad \mu \in M(\mathbb{R}^d), \mu(A) = \mu(-A), \forall A \subset \mathbb{R}^d \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d_+} \cos\big(h\cdot\omega\big)\mu(d\omega), \ \mu \in M(\mathbb{R}^d_+) \end{aligned} \tag{5.4}$$

Therefore, instead of directly defining a distribution on the set of positive definite functions we define a distribution on the measure space $\big(M(\mathbb{R}^d_+), \mathcal{M}(\mathbb{R}^+_d)\big)$ and consider the distribution on the cone of positive definite functions induced by the relation given in (5.4).

5.2.1 Lévy Random Measures: Priors on spectral measures

A simple prior distribution for the spectral measure can be constructed as follows: Fix real numbers $\gamma, \alpha, a, b > 0$, and let G_0 be a probability measure on $\mathbb{R}^d_+$. Let

$\gamma \sim \text{Ga}(a, b)$, and set

$$\mathcal{L} \sim \gamma \text{DP}(\alpha G_0) \tag{5.5}$$

where $\text{DP}(\alpha G_0)$ denotes a Dirichlet process (reference Ferguson) with base measure G_0 and precision parameter α.

Linear combinations of point masses are dense in weak star topology (give reference). Thus, eventhough the realizations from a Dirichlet process are almost surely discrete, the random measure $\mathcal{L}$ constructed above has full support on $M(\mathbb{R}^+_d)$. More generally, for any positive sigma-finite measure $\nu(du, d\omega)$ on $\mathbb{R} \times \mathbb{R}^d_+$ satisfying the integrability condition

$$\iint_{\mathbb{R}\times\mathbb{R}^d_+} (1 \wedge |u|)\nu(du, d\omega) < \infty \tag{5.6}$$

the Lévy random measure

$$\mathcal{L} \sim \text{Lévy}(\nu) \tag{5.7}$$

is a natural choice for a prior distribution for the spectral measure μ.

5.3 Representations of covariance functions: Examples

Example 5.3.1. *(Bochner's theorem continued) Let $d = 1$ and $\lambda(d\omega)$ be a nonnegative symmetric measure on $\mathbb{R}$ with finite support,*

$$\lambda(d\omega) \equiv \sum_{j=1}^{J} \lambda_j \delta_{\omega_j}(d\omega) \tag{5.8}$$

for some $J \in \mathbb{N}$, $\lambda_j \in \mathbb{R}_+$. Define

$$C(h) \equiv \int_{\mathbb{R}} e^{-ih\omega} \lambda(d\omega), \ h \in \mathbb{R}. \tag{5.9}$$

It follows that, for any $N \in \mathbb{N}$, $z \in \mathbb{C}^N$ and $s \in \mathbb{R}^N$,

$$\begin{aligned}\sum_{n,m=1}^{N,N} z_n C(s_n - s_m)\bar{z}_m &= \sum_{n,m=1}^{N,N} z_n \int_{\mathbb{R}} e^{-i(s_n - s_m)\omega} \lambda(d\omega)\bar{z}_m \\ &= \Big|\sum_{n=1}^{N}\sum_{j=1}^{J} z_n e^{-is_n\omega_j}\sqrt{\lambda_j}\Big|^2 \geq 0.\end{aligned}$$

However, as can be seen from the above equation, the quadratic form $\sum_{n,m=1}^{N,N} z_n C(s_n - s_m)\bar{z}_m$ is zero if the N-vector $z \in \mathbb{C}^N$ is orthogonal to each of the J N-vectors $\omega^{(j)} \equiv \{e^{-is_n\omega_j}\sqrt{\lambda_j}\} \in \mathbb{C}^N$, which is always possible if $N > J$. The nonnegative-definite covariance matrix $\Sigma_{N\times N}[n,m] \equiv C(s_n - s_m)$ can be written as

$$\Sigma_{N\times N}[n,m] = \sum_{j=1}^{J} e^{-i(s_n - s_m)\omega_j}\lambda_j = U\Lambda U^*,\ 1 \leq n,m \leq N, \tag{5.10}$$

with $U_{N\times J}[n,j] \equiv e^{-is_n\omega_j}$, $\Lambda_{J\times J}[j,k] \equiv \lambda_j 1_{\{j=k\}}$ and rank(U) = min(N,J). Let $\nu(du,d\omega)$ be a finite measure on $\mathbb{R}_+ \times \mathbb{R}_+$. If we set

$$C(h) = \int_{\mathbb{R}_+} e^{-ih\omega}\mathcal{L}(d\omega),\ \mathcal{L}(d\omega) \sim \textit{Lévy}\big(\nu\big), \tag{5.11}$$

the above discussion implies that the random measure $\mathcal{L}$ must have atleast N number of points in its support to obtain a nonsingular covariance matrix in $\mathbb{R}^N$, where N is the number of locations.

Example 5.3.2. *(Isotropic covariance functions): An arbitrary isotropic covariance function in $\mathbb{R}^d$ can be expressed as [Yaglom 1987, section 22]*

$$C(h) = \int_0^\infty \left(\frac{r|h|}{2}\right)^{-\kappa} \Gamma(\kappa+1) J_\kappa(r|h|)\,\gamma(dr),\ \kappa \equiv \frac{d}{2} - 1, \tag{5.12}$$

where $\gamma(dr)$ is a finite positive measure on $\mathbb{R}_+$, $\Gamma(\cdot)$ is the Gamma function, and

$J_\kappa(\cdot)$ is the Bessel function of order κ defined by (Abramowitz and Stegun [1974])

$$J_\kappa(z) \equiv \frac{(z/2)^\kappa}{\sqrt{\pi}\,\Gamma(\kappa+1/2)} \int_0^\pi \cos(z\cos\theta)(\sin\theta)^{2\kappa} d\theta. \tag{5.13}$$

Hence, as in the previous example, the measure $\gamma(dr)$ can be given a Lévy $\big(\nu\big)$ prior.

Even if the possible singularity of the covariance matrices sampled by the prior can be tackled by incorporating nugget effects (measurement errors), using linear combinations of delta measures may not lead to sparse representations of spectral measures. Finite linear combinations of delta measures as prior distributions for the spectral measure will lead to efficient estimation only if the spectral measure is compactly supported, *i.e.*, the covariance function from which the data is generated is smooth (give reference). Since the spectral measure may not be compactly supported, we seek for more efficient priors. A natural alternative is to first specify a Lévy random field prior for the spectral density $f(\omega)$, and use (5.4) to obtain the induced prior distribution on the covariance function $C(\cdot)$.

Let Θ be a Polish space, and $K : \mathbb{R}^d \times \Theta :\to \mathbb{R}_+$ be a continous function. Define

$$f(\omega) \equiv \int_\Theta K(\omega,\theta)\mathcal{L}(d\theta),\ \mathcal{L} \sim \text{Lévy}(\nu). \tag{5.14}$$

By (5.4), and assuming that change of order of integration is allowed,

$$\begin{aligned} C(h) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp\big(-ih\cdot\omega\big)\mu(d\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp\big(-ih\cdot\omega\big)\, f(\omega) d\omega \\ &= \frac{1}{(2\pi)^{d/2}} \iint_{\mathbb{R}^d\times\Theta} \exp\big(-ih\cdot\omega\big)\, K(\omega,\theta)\, \mathcal{L}(d\theta)\, d\omega. \end{aligned} \tag{5.15}$$

The mixture prior for the spectral density in (5.14), in contrast to the point mass prior, simultaneously captures multiple frequencies with each component in the mixture. Hence such a mixture prior can well approximate a spectral density

with relatively fewer terms and they also sample efficiently from the cone of positive definite functions. For some choices of the function $K(\cdot,\cdot)$, the inner integral in (5.15) has a closed form. We present a few examples.

Example 5.3.3 (Mixture of Gaussian kernels)**.** *Let ν be a measure on the set of positive definite matrices of order d. Consider the scale mixture of d-variate Gaussian kernels with mean 0 and variance Σ, i.e.,*

$$f(\omega) = \int \exp\left\{-\frac{1}{2}\omega'\,\Sigma\,\omega\right\}\mathcal{L}(d\Sigma),\ \mathcal{L} \sim \textit{Lévy}(\nu). \tag{5.16}$$

In this case the covariance function can be expressed as

$$C(h) = \int \exp\left\{-\frac{1}{2}h'\,\tilde{\Sigma}\,h\right\}\mathcal{L}\,(d\tilde{\Sigma}),\ \mathcal{L} \sim \textit{Lévy}(\nu). \tag{5.17}$$

Therefore, a scale mixture of centered Gaussian kernels as a prior distribution for the spectral density yields a mixture of (anisotropic) centered Gaussian covariance functions as the prior distribution for $C(h)$. Since the tails of a centered Gaussian decay rapidly to zero, they are efficient only if the spectral density of the data also decreases to zero rapidly. In particular, priors given by (5.17) *do not well represent non-smooth covariance functions. Gaussian kernels which are not centered can be used in this example to model spectral densities which are heavy tailed, however the induced covariance functions do not have a closed form.*

Example 5.3.4. *(Mixture of Materns) Let $K(\omega|\Sigma,\zeta) \equiv (1+\omega'\Sigma^{-1}\omega)^{-(\zeta+d/2)}$ such that,*

$$f(\omega) = \int K(\omega|\Sigma,\zeta)\mathcal{L}(d\Sigma,d\zeta),\ \mathcal{L} \sim \textit{Lévy}(\nu) \tag{5.18}$$

where $\nu(d\Sigma,d\zeta)$ is a Lévy measure. In this case exchanging the order of integration yields a covariance function that is a mixture of anisotropic Matern covariance

functions,

$$C(h) = \int \frac{\sigma^2}{2^{\zeta-1}\Gamma(\zeta)} \left[\zeta\, h'\Sigma^{-1}h\right]^{\zeta/2} \mathcal{K}_\zeta\left(\left[\zeta h'\Sigma^{-1}h\right]^{1/2}\right) \mathcal{L}(d\Sigma, d\zeta) \tag{5.19}$$

where $\mathcal{K}_\zeta$ is the modified Bessel function of the second kind of order ζ ([Abramowitz and Stegun 1974, section 9.6]). In the case of t-kernels ($\Sigma = \rho\mathbf{I}, \rho \in \mathbb{R}_+$) we recover the construction in Ecker and Gelfand [1999] for isotropic covariance functions.

Example 5.3.5. *[A dense model for covariance function on $\mathbb{R}$] Let ν be a measure on $\mathbb{R}_+ \times \mathbb{R}_+$. Consider the location-scale mixture of exponential kernels as a prior distribution for spectral density,*

$$g(\omega) = \iint_{\mathbb{R}_+\times\mathbb{R}_+} \exp\left\{-\frac{\omega-\mu}{\sigma}\right\} \mathbf{1}_{\omega>\mu} \mathcal{L}(d\mu, d\sigma),\ \mathcal{L} \sim \textit{Lévy}(\nu).$$

This prior distribution has full weak star support on the set of finite positive measures on $\mathbb{R}_+$. By exchanging the order of integration it follows that the induced prior distribution on the covariance function has the form

$$C(h) = \int \left[\frac{\sigma}{1+\sigma^2h^2}\cos(\mu h) - \frac{\sigma^2 h}{1+\sigma^2h^2}\sin(\mu h)\right] \mathcal{L}(d\mu, d\sigma) \tag{5.20}$$

Figure 5.3 shows plots of the integrand for $\sigma = 1$ and different values of μ.

5.4 Estimation: Gaussian Process

Let $Z(s),\ s \in S \subset \mathbb{R}$ be a stationary mean 0 Gaussian process and we observe a realization of Z at a set of locations $\{s_n\}_{n=1}^N$. Our goal in this section is to estimate the covariance function of Z,

$$C(s_n - s_m) = \mathbb{E}\big(Z(s_n) - Z(s_m)\big).$$

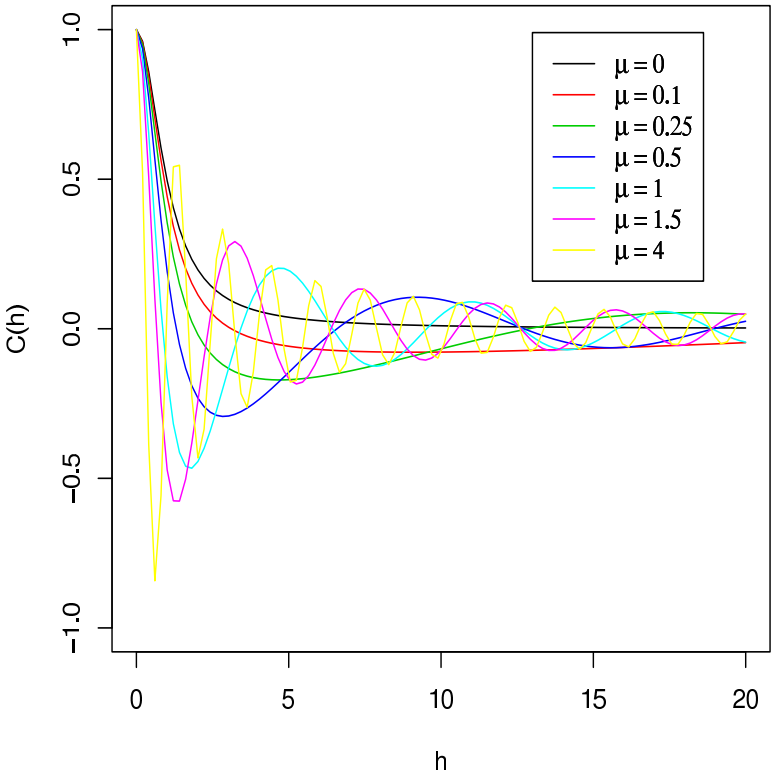

Figure 5.1: Covariance functions from an exponential mixture model

The likelihood of the observed data $Z \equiv (Z(s_1), Z(s_2), \cdots, Z(s_N))$ (with a "nugget effect" τ) is given by

$$L(Z|\ C, \tau) \equiv \frac{1}{(2\pi)^{N/2}|\Sigma|^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2}\ Z'(\Sigma + \tau^2 I)^{-1} Z \Big\} \tag{5.21}$$

where $\Sigma_{N\times N}[m, n] = C(s_n - s_m)$.

In the spatial setting large sample properties of estimators (Fisher Information, MLE, posterior mean *etc.*) can be studied under different conditions (see [Xia 2006]). We discuss two situations: the outfill and the infill asymptotics.

1. Outfill : In the outfill case (expansion asymptotics), the euclidean distance between two neighboring locations remains constant while the domain S is necessarily unbounded. In this case it is known that, under mild regularity conditions, the MLEs are consistent and asymptotically normal.

2. Infill: In the infill case, the domain S is bounded but the distance between the observed locations goes to zero and eventually we obtain every where dense samples inside S. As noted in [Xia 2006, Ying 1991], in the infill case it is not always possible to estimate the parameters of a covariance function consistently even with an infinite amount of data. When the domain S remain fixed and the locations are sampled increasingly dense in S, the observed data points can be highly correlated. For covariance functions from the standard parametric families such as exponential or Matern, the fisher information can be shown to be bounded above ([Xia 2006, pg. 64]). For *e.g.*, for the standard exponential covariance function,

$$C(s,t) \equiv \sigma^2 \exp(-\phi|s-t|), \quad s,t \in S$$

Ying [1991] shows that σ^2, ϕ cannot be estimated consistently, but the MLE for the product $\sigma^2\phi$ is consistent and asymptotically normal with the usual $\sqrt{n}$ rate.

Therefore the effect of prior distribution will be significant even in abundance of data in the infill case. The idea of equivalent Gaussian measures (see [Stein 1999, chapter 4]) is useful in the infill case for prediction even if the parameters cannot be estimated consistently.

5.5 Simulation Examples

We present two examples.

1. A centered Gaussian process was simulated at $N = 250$ equally spaced points in the interval $[1, 100]$ with the covariance function $C(h) = 5\exp(-2\,h^2)$ and with a nugget variance $\tau^2 = 0.01$. The covariance function was given the Gaussian mixture

prior discussed in example 5.3.3:

$$\begin{aligned} C(h) &= \int_{\mathbb{R}_+} \exp\left(-\omega\, h^2\right)\, \mathcal{L}(d\omega) \\ \mathcal{L}(d\omega) &= \text{Lévy}(\nu) \\ \nu(du, d\omega) &= \nu_+ \frac{1}{b^a\, \Gamma(b)} u^{a-1} \exp\left(-u/b\right) \frac{1}{v^m\, \Gamma(v)} \omega^{m-1} \exp\left(-\omega/v\right) \end{aligned}$$

where $\nu_+, a, b, m, v \in \mathbb{R}_+$.

2. A centered Gaussian process was simulated at $N = 250$ equally spaced points in the interval $[1, 100]$ with the covariance function $C(h) = 3\exp\left(-2\, h^2\right)$ and with a nugget variance $\tau^2 = 0.01$. The covariance function was given the mixture prior discussed in example 5.3.5:

$$\begin{aligned} C(h) &= \iint_{\mathbb{R}_+ \times \mathbb{R}_+} \left[\frac{\sigma}{1+\sigma^2 h^2} \cos(\mu h) - \frac{\sigma^2 h}{1+\sigma^2 h^2} \sin(\mu h)\right] \mathcal{L}(d\mu, d\sigma) \\ \mathcal{L}(d\mu, d\sigma) &= \text{Lévy}(\nu) \\ \nu(du, d\mu, d\sigma) &= \nu_+ \frac{1}{b^a\, \Gamma(b)} u^{a-1} \exp\left(-u/b\right) \frac{1}{v^m\, \Gamma(v)} \mu^{m-1} \exp\left(-\mu/v\right) \frac{1}{s^m\, \Gamma(s)} \sigma^{m-1} \exp\left(-\sigma/v\right) \end{aligned}$$

where $\nu_+, a, b, m, v, s \in \mathbb{R}_+$.

The simulations were performed using a reversible jump MCMC algorithm. Figure 5.2 shows the draws from the posterior distribution for the first simulation. In this case the kernel was chosen from the same parametric family as the true covariance distribution, and hence as expected, the posterior mean estimates the true covariance function very well and the posterior quantiles are very tight. The mean number of terms in the stochastic representation of the covariance function $C(h)$ was 1.2. This shows that if there is a sparse representation possible in the overcomplete dictionary expansion, our model will capture it.

Figure 5.3 shows the results from the second simulation study. In this case the kernels do not have the same parametric form as the true covariance function. How-

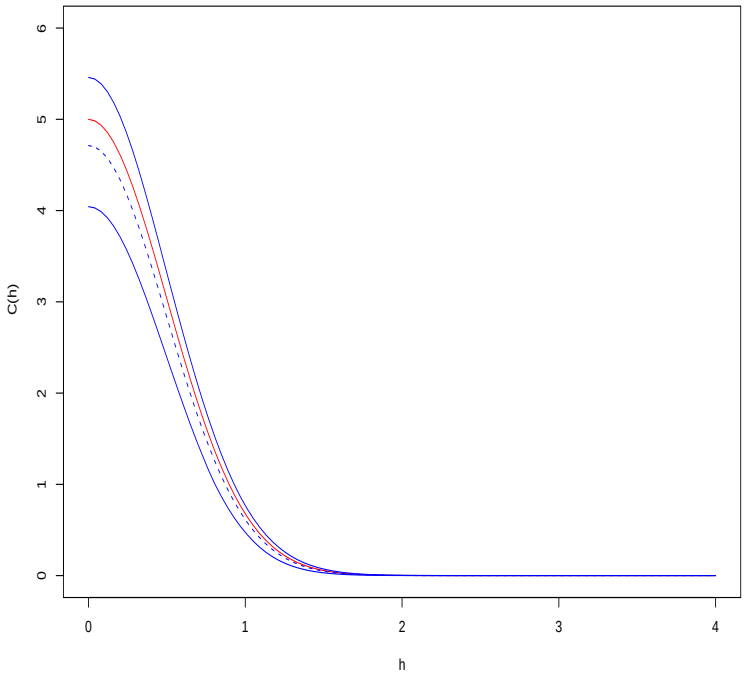

Figure 5.2: The red curve is true covariance function, the dotted blue curve is posterior mean, the blue curves are the 5% and 95% quantiles from the posterior draws.

ever the posterior mean estimates the covariance function fairly well (even though not as good as the previous case) and the true function is contained in the quantile curves. The mean number of terms in the stochastic representation of the covariance function $C(h)$ was 4.1. Both the above simulations were performed for different values of hyper parameters and the final results were not sensitive to their values.

5.6 Discussion

We have illustrated a novel application of Lévy random measures for estimating of covariance functions of stationary stochastic processes. Several examples were discussed, and simulation studies were carried out in a one dimensional example.

The one dimensional simulation results are encouraging. However in higher di-

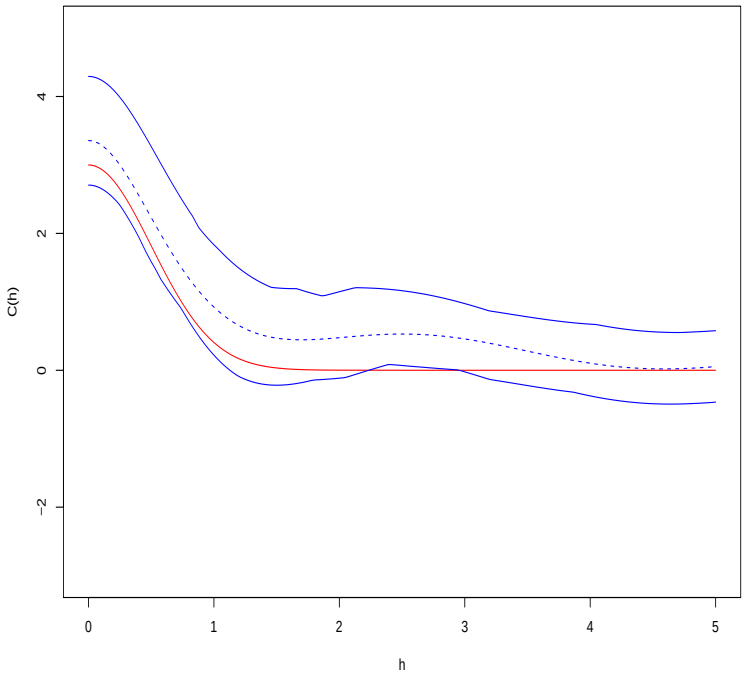

Figure 5.3: The red curve is true covariance function, the dotted blue curve is posterior mean, the blue curves are the 5% and 95% quantiles from the posterior draws.

mensions, for efficient estimation more care needs to be given in specifying the prior distribution, especially in the choice of kernels. Furthermore, results regarding posterior consistency must be obtained to ensure the correctness of the methodology.

Bibliography

Abramowitz, M. and Stegun, I. A., eds. (1974). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. Dover, New York, NY.

Albiac, F. and Kalton, N. J. (2006). *Topics in Banach space theory*, *Graduate Texts in Mathematics*, vol. 233. Springer-Verlag, New York, NY.

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli* **9** 291–312.

Applebaum, D. (2004). *Lévy processes and Stochastic Calculus*, *Cambridge Studies in Advanced Mathematics*, vol. 93. Cambridge Univ. Press, Cambridge, UK.

Aronszajn, N. (1950). Theory of reproducing kernels. *T. Am. Math. Soc.* **686** 337–404.

Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Stat.* **27** 536–561.

Baum, L. E., Katz, M. and Read, R. R. (1962). Exponential convergence rates for the law of large numbers. *Trans. Amer. Math. Soc.* **102** 187–199.

Belkin, M. and Niyogi, P. (2004). Semi-supervised learning on Riemannian manifolds. *Machine Learning* **56** 209–239.

Belkin, M., Niyogi, P. and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* **7** 2399–2434.

Billingsley, P. J. (1999). *Convergence of Probability Measures*. Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, New York, NY, second edn.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1** 121–143 (electronic).

Bochner, S. (1955). *Harmonic analysis and the theory of probability*. University of California Press, Berkeley and Los Angeles.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *J. Mach. Learn. Res.* **2** 499–526.

Carter, C. K. and Kohn, R. (1997). Semiparametric Bayesian inference for time series with mixed spectra. *J. Roy. Statist. Soc. Ser. B* **59**.

Chakraborty, S., Ghosh, M. and Mallick, B. K. (2005). Bayesian nonlinear regression for large p small n problems. *J. Am. Stat. Assoc.* Under revision.

Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *J. Multivariate Anal.* **98** 1969–1987.

Choudhuri, N., Ghosal, S. and Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *J. Am. Stat. Assoc.* **99** 1050–1059.

Clyde, M. A. and Wolpert, R. L. (2007). Estimation using oveercomplete dictionaries. In *Bayesian Statistics 8* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.), 91–114. Oxford University Press, Oxford, UK.

Cortes, C. and Vapnik, V. N. (1995). Support-Vector Networks. *Machine Learning* **20** 273–297.

Cramér, H. (1940). On the theory of stationary random processes. *Ann. of Math. (2)* **41** 215–230.

Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society* **39** 1–49.

de Boor, C. and Lynch, R. E. (1966). On splines and their minimum properties. *J. Math. Mech.* **15** 953–969.

DeVore, R. A., Howard, R. and Micchelli, C. A. (1989). Optimal nonlinear approximation. *Manuskripta Mathematika* .

Diaconis, P. (1988). Bayesian numerical analysis. In *Statistical decision theory and related topics, IV* (S. S. Gupta and J. O. Berger, eds.), vol. 1, 163–175. Springer-Verlag, New York, NY.

Diaconis, P. and Freedman, D. A. (1986a). On inconsistent Bayes estimates of location. *Ann. Stat.* **14** 68–87.

Diaconis, P. and Freedman, D. A. (1986b). On the consistency of Bayes estimates. *Ann. Stat.* **14** 1–67.

Dunford, N. and Schwartz, J. T. (1988). *Linear operators. Part I.* Wiley Classics Library, John Wiley & Sons, New York, NY.

Ecker, M. D. and Gelfand, A. E. (1999). Bayesian modeling and inference for geometrically anisotropic spatial data. *Mathematical Geology* **31** 67–83.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.* **90** 577–588.

Evgeniou, T., Pontil, M. and Poggio, T. (2000). Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics* **13** 1–50.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1** 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Stat.* **2** 615–629.

Fredholm, E. I. (1900). Sur une nouvelle méthode pour la résolution du problèm de Dirichlet. *Euvres complètes:publiées sous les auspices de la Kungliga svenska vetensakademien par l'Institut Mittag-Leffler* 61–68.

Gaigalas, R. (2004a). A Poisson bridge between fractional Brownian motion and stable Lévy motion. Tech. rep., Department of Mathematics, Uppsala University.

Gaigalas, R. (2004b). The "polish" integral with respect to compensated Poisson random measures. Personal communication.

Gangopadhyay, A. K., Mallick, B. K. and Denison, D. G. T. (1999). Estimation of spectral density of a stationary time series via an asymptotic representation of the periodogram. *J. Statist. Plann. Inference* **75** 281–290. The Seventh Eugene Lukacs Conference (Bowling Green, OH, 1997).

Ghosal, S. and Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *Ann. Stat.* **34** 2413–2429.

Ghosal, S. and van der Vaart, A. W. (2007a). Convergence rates of posterior distributions for noniid observations. *Ann. Stat.* **35** 192–223.

Ghosal, S. and van der Vaart, A. W. (2007b). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Stat.* **35** 697–723.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin* 49–52.

Hájek, J. (1961). On a property of normal distributions of any stochastic process. *Select. Transl. Math. Statist. and Probability* **1** 245–252.

HÁJEK, J. (1962). On linear statistical problems in stochastic processes. *Czechoslovak Math. J.* **12** 404–444.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag.

ITI SATO, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*, *Cambridge Studies in Advanced Mathematics*, vol. 68. Cambridge Univ. Press, Cambridge, UK.

JAMES, L. F., LIJOY, A. and PRÜNSTER, I. (2005). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Stat.* **33** 105–120.

JOHNSTONE, I. (1998). Function estimation in Gaussian noise: sequence models. Draft of a monograph.

KALLIANPUR, G. (1970). The role of reproducing kernel Hilbert spaces in the study of Gaussian processes. *Advances in Probability and Related Topics* **2** 49–83.

KÖNIG, H. (1986). *Eigenvalue distribution of compact operators*, *Operator Theory: Advances and Applications*, vol. 16. Birkhäuser, Basel, CH.

KIMELDORF, G. S. and WAHBA, G. (1971). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.* **41** 495–502.

LIANG, F., MAO, K., LIAO, M., MUKHERJEE, S. and WEST, M. (2007a). Non-parametric Bayesian kernel models. Discussion Paper 2007-10, Duke University ISDS, Durham, NC.

LIANG, F., MUKHERJEE, S. and WEST, M. (2007b). The use of unlabelled data in predictive modeling. *Stat. Sci.* **22** 189–205.

LUKIĆ, M. N. and BEDER, J. H. (2001). Stochasitic processes with sample paths in reproducing kernel Hilbert spaces. *T. Am. Math. Soc.* **353** 3945–3969.

MACEACHERN, S. and MÜLLER, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Stat.* 223–238.

MAZJA, V. G. (1985). *Sobolev Spaces*. Springer-Verlag, New York, NY.

MÜLLER, P., QUINTANA, F. and ROSNER, G. (2004). A method for combining inference across related nonparametric Bayesian models. *J. Am. Stat. Assoc.* 735–749.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London A* **209** 415–446.

Micchelli, C. A. and Wahba, G. (1981). Design problems for optimal surface interpolation. In *Approximation Theory and Applications* (Z. Ziegler, ed.), 329–348.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R. M., Engle, A., Campbell, C., Golub, T. R. and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA Microarray data. *Journal of Computational Biology* **10** 119–143.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer, New York. Lecture Notes in Statistics 118.

Nieto-Barajas, L. E., Prünster, I. and Walker, S. G. (2004). Normalized random measures driven by increasing additive processes. *Ann. Stat.* **32** 2343–2360.

Ouyang, Z., Clyde, M. A. and Wolpert, R. L. (2008). Nonparametric Bayesian kernel regression and classification with feature selection. Discussion Paper (in preparation), Duke University Statistical Science, Durham, NC.

Parzen, E. (1963). Probability density functionals and reproducing kernel Hilbert spaces. In *Proceedings of the Symposium on Time Series Analysis* (M. Rosenblatt, ed.), 155–169. John Wiley & Sons, New York, NY.

Pawitan, Y. and O'Sullivan, F. (1994). Nonparametric spectral density estimation using penalized Whittle likelihood. *J. Amer. Statist. Assoc.* **89** 600–610.

Poggio, T. and Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science* **247** 978–982.

Poggio, T., Rifkin, R. M., Mukherjee, S. and Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature* **428** 419–422.

Rajput, B. S. and Rosiński, J. (1989). Spectral representations of infinitely divisible processes. *Probab. Theory Rel.* **82** 451–487.

Ramaswamy, S., Tamayo, P., Rifkin, R. M., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S. and Golub, T. R. (2001). Multiclass cancer diagnosis using tumor gene expression signatures.

Proc. Nat. Aca. Sci. **98** 149–54.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

Reed, M. C. and Simon, B. (1980). *Methods of Modern Mathematical Physics. I. Functional Analysis*. Academic Press, New York, NY, second edn.

Reich, B. J. and Fuentes, M. (2007). A multivariate semiparametric bayesian spatial modeling framework for hurricane surface wind fields. *Annals of Applied Statistics* **1** 249–264.

Rogers, L. C. G. and Williams, D. (1987). *Diffusions, Markov Processes, and Martingales*, vol. 2. John Wiley & Sons, New York, NY.

Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Schoenberg, I. J. (1942). Positive definite functions on spheres. *Duke Mathematics Journal* **9** 96–108.

Shawe-Taylor, J. S. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, Cambridge, UK.

Sollich, P. (2002). Bayesian methods for Support Vector Machines: Evidence and predictive class probabilities. *Machine Learning* **46** 21–52.

Stein, M. L. (1999). *Interpolation of spatial data*. Springer Series in Statistics, Springer-Verlag, New York. Some theory for Kriging.

Taqqu, M. S. and Wolpert, R. L. (1983). Infinite variance self-similar processes subordinate to a Poisson measure. *Z. Wahrscheinlichkeit.* **62** 53–72.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Doklady* **4** 1035–1038.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1** 211–244.

Tokdar, S. T. and Ghosh, J. K. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *J. Statist. Plann. Inference* **137** 34–42.

Tu, C., Clyde, M. A. and Wolpert, R. L. (2006). Lévy adaptive regression

kernels. Discussion Paper 2006-08, Duke University Department of Statistical Science.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics, Springer-Verlag, New York, NY.

VAPNIK, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons, New York, NY.

WAHBA, G. (1990). *Splines Models for Observational Data*, *Series in Applied Mathematics*, vol. 59. SIAM, Philadelphia, PA.

WAHBA, G. (1999). Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In *Advances in Kernel Methods: Support Vector Learning* (B. Schölkopf, A. J. Smola, C. J. C. Burges and R. Soentpiet, eds.), 69–88. MIT Press, Cambridge, MA.

WASSERMAN, L. (2005). *All of Nonparametric Statistics*. Springer-Verlag.

WEHAUSEN, J. V. (1938). Transformations in linear topological spaces. *Duke Math. J.* **4** 157–169.

WEST, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. Discussion Paper 1992-03, Duke University ISDS, Durham, NC.

WOLPERT, R. L. and ICKSTADT, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85** 251–267.

WOLPERT., R. L. and ICKSTADT, K. (1998). Simulation of Lévy random fields. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. K. Dey, P. Müller and D. Sinha, eds.), *Lecture Notes in Statistics*, vol. 133, 305–322. Springer-Verlag, New York, NY.

WOLPERT, R. L. and ICKSTADT, K. (2004). Reflecting uncertainty in inverse problems: A Bayesian solution using Lévy processes. *Inverse Problems* **20** 1759–1771.

WOLPERT, R. L., ICKSTADT, K. and HANSEN, M. B. (2003). A nonparametric Bayesian approach to inverse problems. In *Proceedings of the 7th Valencia International Meeting held in Tenerife, June 2–6, 2002* (J. M. Bernardo, A. P. Dawid, J. O. Berger, M. West, D. Heckerman, M. J. Bayarri and A. F. M. Smith, eds.), 403–417. Cambridge Univ. Press, New York, NY. With a discussion by Subhashis Ghosal and a rejoinder by the authors.

Wolpert, R. L. and Taqqu, M. S. (2005). Fractional Ornstein-Uhlenbeck Lévy processes and the Telecom process: Upstairs and downstairs. *Signal Processing* **85** 1523–1545.

Xia, G. (2006). On large sample issues in spatial statistics. Ph.D. thesis, Department of Statistical Science, Duke University, Durham, NC.

Xing, E. P., Sharan, R. and Jordan, M. I. (2004). Bayesian haplotype inference via the Dirichlet process. In *Machine Learning, Proceedings of the 21st International Conference (ICML 2004), Banff, Canada* (C. E. Brodley, ed.). ACM Press, New York, NY.

Xing, E. P., Sohn, K.-A., Jordan, M. I. and Teh, Y.-W. (2006). Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture. In *Machine Learning, Proceedings of the 23rd International Conference (ICML 2006), Pittsburgh, PA* (W. Cohen and A. Moore, eds.). ACM Press, New York, NY.

Yaglom, A. M. (1987). *Correlation theory of stationary and related random functions. Vol. I.* Springer Series in Statistics, Springer-Verlag, New York. Basic results.

Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *J. Multivariate Anal.* **36** 280–296.

Zhou, D.-X. (2003). Capacity of reproducing kernel spaces in learning theory. *IEEE T. Inform. Theory* **49** 1743–1752.

Biography

Natesh S. Pillai was born in 1981, at Thiruvanathapuram in southern India. He obtained his Bachelor's degree in 2003, from Indian Institute of technology, Chennai. He obtained his Master's degree from the Department of Statstical Sciences, Duke University in 2007. He enjoys a variety of things including classical music, chess, social service and teaching.