Bayesian inference for Gaussian models: Inverse problems and evolution equations

Yan, D.

Citation

Yan, D. (2020, March 3). *Bayesian inference for Gaussian models: Inverse problems and evolution equations*. Retrieved from https://hdl.handle.net/1887/86070

Author: Yan, D.
Title: Bayesian inference for Gaussian models: Inverse problems and evolution equations
Issue Date: 2020-03-03

Chapter 1

Introduction

The objective of statistical inference is to infer a quantity from observations. A 'useful' inference procedure should disclose a certain truth. Intuitively, it sounds reasonable to wish that, given the existence of a (fixed) underlying truth, the inference procedure produces approximations closer to the truth, when more information is available. Following this intuition, a rigorous theory has been well developed in the twentieth century, which is now known as the field of *asymptotic theory* (alternatively, *large sample theory*). To reflect realistic observational processes, it is customary to number observations by the natural number $\mathbb{N} = \{1, 2, 3, \cdots\}$. In the asymptotic framework, the performance of inferences is studied using various criteria measuring the 'accuracy' of an estimation, when the sample size n tends to infinity. The behaviour of an inferential procedure by taking the limit as $n \to \infty$ is called *asymptotics*.

The idea in the previous paragraph can be formulated in mathematical language as follows. Given a measurable space $(\mathbb{X}, \mathscr{X})$, i.e. the *sample space*, an *experiment* $\mathscr{E}$ is a set $\{\mathbb{P}_\theta : \theta \in \Theta\}$ of probability distributions indexed by a parameter family Θ. In practice, the indexation is guided by a model that generates a probability distribution $\{\mathbb{P}_\theta\}$ for each $\theta \in \Theta$. In this situation, the experiment is also called a *statistical model*. For each $n \in \mathbb{N}$, the nth observation $X^{(n)}$ is a random element whose distribution $\mathbb{P}_\theta^{(n)}$ is from an experiment $\mathscr{E}_n = \{\mathbb{P}_\theta^{(n)}\}_{\theta \in \Theta}$. The term *sample* and *observation* are used interchangeably for $\{X^{(n)}\}$. Notice that the experiments are not necessarily identical (as they depend on n), but they are indexed with the same parameter space, and additionally, each sample $X^{(n)}$ encodes some information on the same parameter θ. Statistical inference is to propose an estimate[1] $\widehat{\theta}$, which only depends on the observations, that is a sensible approximation to the true parameter θ.

In the general asymptotic framework, the following components are decisive to the formulation of asymptotics: the parameter space Θ, the experiment $\mathscr{E}_n$, and the methodology guiding the inferential process. In this thesis, we will study the *asymptotics* of *Bayesian nonparametric* inference for *Gaussian linear models*

---

[1] In literature, statistical inference is often categorised into estimation and hypothesis testing. However, we use inference and estimation interchangeably, as hypothesis testing is not touched in this thesis.

(GLMs). Admittedly confusing, the term 'Nonparametric' actually refers to the fact that the dimension of the parameter space Θ is infinite. In order to handle statistical analysis in infinite-dimensional space, a set of mathematical tools, which differs from the tools used in parametric statistics, is desired. We will provide a short survey in Chapter 2. The statistical model considered is the Gaussian linear model (GLM). To be precise, we will study several models, varying in the level of abstraction and generality, all of which fall into the class of GLMs. The phrase 'Bayesian' refers to the methodology that is based on Bayes' rule. In Chapter 3, the general framework of GLMs will be outlined and a Bayesian asymptotic result coping with GLM will also be given.

The rest of the chapter is arranged as follows. First we sketch the characteristics of parameter spaces used in nonparametric statistics in Section 1.1. Then, the statistical model considered in this thesis is introduced in Section 1.2. The Bayesian approach to statistical inference is briefly reviewed in Section 1.3. After specifying all the necessary components to formulate a feasible asymptotic study, we provide in Section 1.4 an overview of the main topics covered in this thesis. We conclude this chapter with a summary of notation in Section 1.5, and supplementary notes with references in Section 1.6.

1.1 Infinite-Dimensional Parameter Spaces

To accommodate complex stochastic phenomena, many advanced probabilistic models have been naturally developed in infinite-dimensional spaces. Meanwhile, due to the growing capacity of computation and the increasing volume of data storage, it is also practical to consider inference for models of infinite-dimensional nature, which leads to the study of nonparametric estimation. The parameter spaces in parametric inference are essentially finite-dimensional vector spaces, i.e. Euclidean spaces. Due to its plain structure, it does not raise any concerns that the chosen space may give rise to complications in the estimation. On the contrary, in nonparametric statistics, the basic assumption is that the true parameter belongs to an infinite-dimensional space Θ, whose structure is more involved. For example, Lebesgue measure does not extend to infinite-dimensional spaces; closed and bounded sets are not necessarily compact, etc. Therefore, infinite-dimensional spaces require more careful examination. In this section, we will introduce the fundamental concept of smoothness of Hölder and Sobolev types. Other related notions separability, approximation property and compactness are also important criteria of nonparametric parameter spaces. A detailed discussion with the emphasis on Hilbert spaces can be found in Chapter 2.

In nonparametric statistics, the infinite-dimensional parameter space Θ is often postulated to be a complete space, i.e. a Banach space. Sometimes, it is also hypothesised that the space Θ possesses an inner product structure, and hence Θ is a Hilbert space. The Hilbertian assumption is often driven by underlying (physical) models, such as the postulates of quantum mechanics. We consider the following two canonical examples of parameter spaces: on $\mathfrak{T} = [0, 1] \subset \mathbb{R}$, the space $C[0, 1]$ of continuous functions, and the space $L^2[0, 1]$ of square integrable functions, which has an inner product structure.

With the index n indicating the growth of information, for a sequence $\{\widehat{\theta}_n\}$ of estimates for the parameter θ, it is desirable that $d(\widehat{\theta}_n, \theta)$ converges to zero in probability as n tends to infinity. If the convergence is true for all possible parameters, then the statistical procedure is (asymptotically) *consistent*. Without further assumptions, the convergence rate may be arbitrarily slow, which is not so useful in practice.

To obtain a reasonable convergence rate, e.g. the polynomial rate $n^{-\beta}$ with $\beta \in \mathbb{R}^+$, the logarithmic rate $\log n$, etc., a *smoothness* (or *regularity*) condition is required. Briefly, a set $\{\Theta_s : s \in \mathcal{S}\}$ of subspaces of a parameter space Θ is called a *smoothness class*, if a reasonable convergence rate can be obtained for certain statistical models, when the parameter is known to belong to Θ_s. The index s in Θ_s symbolises the 'smoothness'. To illustrate this concept, we use the two previously mentioned examples $C[0,1]$ and $L^2[0,1]$.

Hölder Smoothness

Recall that $C[0,1]$ is the space of continuous functions. For $k \in \mathbb{N}$, let $C^k[0,1]$ be the space of k-times differentiable (in the classical sense) continuous functions and $C^0[0,1] = C[0,1]$. For a non-integer s, define

$$C^s[0,1] := \left\{ f \in C^{\lfloor s \rfloor} : \|f\|_s := \sup_{\substack{x,y \in [0,1]: \\ x \neq y}} \frac{|f^{\lfloor s \rfloor}(x) - f^{\lfloor s \rfloor}(y)|}{|x-y|^{s-\lfloor s \rfloor}} < \infty \right\},$$

where $\lfloor s \rfloor$ is the largest integer strictly smaller than s and $f^{\lfloor s \rfloor}$ is the $\lfloor s \rfloor$-th derivative. The spaces $C^s, s \in \mathbb{R}_0^+$ are *Hölder spaces*, which is the canonical smoothness class for continuous functions.

Continuity plays an important role in the study of inference. The standard estimator for a Hölder class, the *kernel estimator*, is constructed by utilising the continuity property, and consequently the convergence rate also depends on the Hölder smoothness.

Sobolev Smoothness

Sobolev smoothness is directly related to the concept of (weak) *differentiability*. Let $\mathfrak{D}$ be a bounded domain in Euclidean space $\mathbb{R}^d$. With $k \in \mathbb{N}$, the Sobolev spaces $H_k(\mathfrak{D})$, containing the $L^2(\mathfrak{D})$ functions whose L^2-weak derivatives exist up to the order k, were first introduced to study partial differential equations (PDEs). Soon it evolved into an important (sub)field in the theory of function spaces, while maintaining an intimate connection to PDEs. One discovery of great importance is that the integer indexed Sobolev spaces, in many situations, e.g. certain boundary conditions being satisfied, can be identified with the domains of the integer powers of a differential operator that is densely defined, strictly positive, and self-adjoint. Subsequently, *fractional* Sobolev spaces can be defined using spectral theory.

With the standard result that the Fourier basis is the eigenbasis of the second order differential operator on $L^2[0,1]$ with the periodic boundary condition, a spectral argument leads to the well known Statistician's Sobolev spaces as follows. The *projection* mapping from a function f in $L^2[0,1]$ to its Fourier coefficients

$\{f_j\}_{j\in\mathbb{N}}$ is an isometric isomorphism from L^2 to the space ℓ^2 of square summable sequences. In addition, with the singular values $\{j : j \in \mathbb{N}\}$ of the second order differential operator, for $s \in \mathbb{R}_0^+$, the Sobolev norms admits a form of weighted ℓ^2 norms, i.e.

$$\|f\|_s^2 = \sum_{j\in\mathbb{N}} j^{2s} f_j^2,$$

in particular, $\|f\|_0 = \|f\|_{\ell^2}$.

In the example $L^2[0,1]$, the Sobolev smoothness is directly linked to an *unbounded operator* (the second differential operator), which induces a scale of Hilbert spaces with duality relations (see related sections in Chapter 2). Due to the Hilbert space structure, the typical estimators constructed for L^2 are commonly based on projections. Similar to Hölder class, the achieved convergence rate depends on the Sobolev smoothness.

In Hölder smoothness and the example of Sobolev smoothness, the index set is $\mathfrak{T} = [0,1]$ for simplicity. As seen in the general description on Sobolev smoothness, it is often possible to be replaced by a more general setting, e.g. *compact metric spaces*, and in particular, closed and bounded domains in Euclidean spaces. When the domain $\mathfrak{T} \subset \mathbb{R}^d$ but not bounded, e.g. $\mathfrak{T} = [0,\infty)^d$, the spaces defined on $\mathfrak{T}$ become too 'big' to recover the parameter. This difficulty is usually removed by introducing additional properties to the parameter, for example, periodicity or tail conditions.

Although having been mentioned separately in the related paragraphs, we want to stress a fundamental difference between Hölder and Sobolev smoothness. In the Hölder case, the smoothness is characterised by the *local properties of paths* of a function, and the index set may directly influence the properties of functions. In statistics, Hölder smoothness is usually chosen for the recovery of stochastic processes whose index sets have straightforward interpretations, such as time, spatial domains, etc. On the other hand, Sobolev smoothness directly relates to the *duality structure*, as in the previous example via weak differentiability, or in general via an (unbounded) operator. Since probability measures on infinite-dimensional spaces are usually (if not always) characterised via the topological duals, Sobolev smoothness serves as a natural stage for studying statistical problems in infinite-dimensional spaces. In particular, if the underlying space is a Hilbert space, the full scale of the induced dual spaces admits explicit representations, which significantly remedies the technicality involved in the development of the theory. A well known case is the L^2 theory in PDEs. In this thesis, we exclusively focus on the estimation in the smoothness of *Hilbertian Sobolev* type, i.e. Sobolev spaces that are also Hilbert spaces.

1.2 Gaussian Linear Models

Arguably, the Gaussian linear model has occupied a central role since the birth of statistics. Its importance cannot be exaggerated. In this section, we heuristically introduce a general (nonparametric) Gaussian linear model in Hilbert space setting. This model provides a common ground for the (more concrete) models studied in the subsequent chapters. We will systematically examine the model in Chapter 3.

Formally, the model can be represented as a very simple additive model,

$$X^{(n)} = \mathcal{A}^{(n)}\theta + \xi^{(n)}. \tag{1.1}$$

The parameter of interest is the *signal* θ. In practice, however, it is often the case that only a transform of the original signal is observable, because of e.g. the experimental set-up, etc. A bounded operator $\mathcal{A}^{(n)}$, from the parameter space Θ to another Hilbert space $\mathbb{X}$, is used to characterise the *transform* (alternatively, *forward mapping*). The randomness arising in the observational process is modelled by a stochastic noise $\xi^{(n)}$ with a structure of Gaussian type.

The model consists of three components: the signal θ, the *forward mappings* $\mathcal{A}^{(n)}$, and the noise $\xi^{(n)}$. Throughout this thesis, the parameter θ is understood as an infinitely-dimensional object, while the features of $\mathcal{A}^{(n)}$ and ξ depend on observation schemes.

Continuous observation
Continuous observation is often an idealization of experiments. The observation $X^{(n)}$ is 'complete': for example, the entire trajectory of a process, all functionals on the parameter space, etc. Although in most cases the observation scheme is unrealistic, the model is fruitful for gaining insight. In this situation, the forward mapping is set to $\mathcal{A}^{(n)} = \mathcal{A}$, see Section 1.2.1 for more information. The noise structure is more involved, see the subsequent section Section 1.2.2.

Discrete observation
Discrete observation reflects a more realistic situation. We consider the concrete situation that the image $\mathcal{A}(\Theta)$ is contained in a function space defined on a domain $\mathfrak{D}$, and noisy samples of the unknown function $\mathcal{A}f$ at a finite number of locations, called *design points*, in its domain are recorded. The observation $Y^{(n)}$ is a random vector $(Y_i)_{i\leq n}$ in $\mathbb{R}^n$, with the coordinates given by

$$Y_i = \mathcal{A}f(x_i) + z_i, \quad i = 1, \cdots, n,$$

where $(z_i)_{i\leq n}$ is often assumed to be standard Gaussian in $\mathbb{R}^n$. As a consequence, the forward mappings are $\mathcal{A}^{(n)} = \mathcal{E}^{(n)}\mathcal{A}$, where $\mathcal{E}^{(n)}$ is the evaluation operator at (deterministic) design points. Alternatively, discrete observation may also refer to noisy finite-dimensional projections of the signal $\mathcal{A}f$. We do not pursue this direction. In fact, under mild conditions, the discrete observation on design points can be translated to finite-dimensional projections, see the relevant chapter Chapter 8.

We are going to describe the forward mapping and the noise below. After that, we also briefly mention the models that will be investigated in later chapters.

1.2.1 Forward Mapping $\mathcal{A}$

Since our goal is to recover the parameter θ, a procedure to reconstruct θ from the image $\mathcal{A}\theta$ is desired, i.e. to 'invert' the forward operator $\mathcal{A}^{(n)}$. The methods working with noiseless observations (i.e. $\xi^{(n)} = 0$ almost surely) do not necessarily remain valid for noisy observations. In fact, most of them break down and modifications are needed. The recovery of θ depends on the following components: the transform $\mathcal{A}^{(n)}$ and the structure of the noise $\xi^{(n)}$. Heuristically, the idea for the recovery is described as follows. For clarity, the superscript (n) is omitted. The structure of ξ will determine whether the observations realise in the space $\mathbb{X}$ or are actually certain 'generalised' processes over $\mathbb{X}$. An approximate 'inverse'[2] $\mathcal{A}^\dagger$ of $\mathcal{A}$ would help to recover θ. Assuming that $\mathcal{A}^\dagger$ is available, formally applying $\mathcal{A}^\dagger$ to X, we obtain the sum of $\mathcal{A}^\dagger\mathcal{A}\theta$ and $\mathcal{A}^\dagger\xi$. $\mathcal{A}^\dagger$ would be acceptable if $\mathcal{A}^\dagger\mathcal{A}\theta$ is a reasonable approximation of θ, while the spread of $\mathcal{A}^\dagger\xi$ is under control. It is worth noticing that $\mathcal{A}^\dagger$ interplays with ξ. Precisely, if $\mathcal{A}^\dagger$ is linear, the $\mathcal{A}^\dagger\xi$ has the covariance[3] $\mathcal{A}^\dagger\Sigma(\mathcal{A}^\dagger)^*$. A desirable 'inverse' $\mathcal{A}^\dagger$ should not amplify the noise too much. From the heuristics above, we conclude that the combination of the transform and the noise to a large extent determines the inferential procedure.

1.2.2 Gaussian noise

As already mentioned at the beginning of this section, the noise ξ is directly related to the observation scheme adopted. For discrete observation, the standard Gaussian on Euclidean space is well-known and it does not need explanation. On the other hand, for continuous observation, the noise deserves a closer look.

Consider the case that ξ is a Gaussian random element that lives 'around'[4] $\mathbb{X}$. There are two interpretations to characterise a Gaussian element. When the underlying space $\mathbb{X}$ contains the functions on a domain $\mathfrak{D}$, the noise ξ can be treated as a *Gaussian process* on the same domain. Alternatively, ξ can also be viewed as a random element whose law is given by a *Gaussian measure* on $\mathbb{X}$. In many situations, they are equivalent and can be translated from one to the other. Since a large class of the functionals of Wiener process are continuous, it is customary to consider the process version when Hölder smoothness is considered. On the other hand, due to the intrinsic Hilbert structure (i.e. *reproducing kernel Hilbert space*, alias RKHS), Gaussian measure version is often adopted for Sobolev smoothness. More details regarding the features of Gaussian elements are given in Chapter 3.

A Gaussian element ξ in a Hilbert space $\mathbb{X}$ is fully characterised by a mean vector and a covariance operator Σ on $\mathbb{X}$. Similar to real-valued Gaussian variables, the mean and the covariance operator specify, respectively, the location and the 'spread' of the distribution. The noise ξ will always be zero mean in this study. If Σ is an operator of trace class, the noise ξ is a *proper* random element in $\mathbb{X}$, i.e. $\xi \in \mathbb{X}$ almost surely. Otherwise, ξ is a *generalised* random element *over* $\mathbb{X}$, that means, ξ takes values in an extension of $\mathbb{X}$. One noteworthy example of a generalised random element is *white noise*, i.e. $\Sigma = \mathrm{id}$, where id is the identity operator on an infinite-dimensional Hilbert space $\mathbb{X}$.

---

[2] Precisely, it is the *psudoinverse*. This concept is only used in this section for heuristics, and hence we refer to [29] for the detailed treatment.

[3] $\mathcal{A}^*$ is the adjoint of $\mathcal{A}$, see Definition A.1.

[4] As we do not specify if ξ takes values in $\mathbb{X}$ almost surely.

1.2.3 Concrete Models

We will examine two types of Gaussian linear models in this thesis. Although the two types are not completely disjoint from each other and have some overlap, each of them has its own flavour and interest.

Inverse Problems

We only consider the linear case, i.e. $\mathcal{A}$ being a linear operator. When $\mathcal{A}$ has no bounded inverse, which frequently occurs when $\mathcal{A}$ has a 'smoothing' property, the (unbounded) $\mathcal{A}^\dagger$ amplifies the noise ξ. The situation is *ill-posed*, as small perturbations (modelled by stochastic noise) in the observation result in a poor reconstruction, dominated by the noise $\mathcal{A}^\dagger\xi$ magnified by the unbounded $\mathcal{A}^\dagger$. As a consequence, *regularization* techniques need to be developed to handle the ill-posedness of $\mathcal{A}^\dagger$. The models satisfying the description above are customarily considered *inverse problems*, and the ill-posedness is also known as *inverse nature.* To focus on examining the ramifications of ill-posedness, the noise structure is selected to be a simple form, white noise, which has been widely accepted as a reasonable choice for nonparametric models, see e.g. [36].

The inverse problem is the main theme of Part II.

Evolution Equations

Stochastic partial differential equations (SPDEs) are broadly applied to model stochastic (differential) dynamical systems, time dependent processes whose dynamics possess differential structures. Due to the involvement of time, they are also known as *evolution equations*. A typical form of linear evolution equations is

$$\begin{cases} dX(t) = [\mathcal{L}X(t) + f(t)]\,dt + B\,dW(t), \\ X(0) = u \end{cases} \tag{1.2}$$

where $\mathcal{L}$ is a differential operator, f is a *source term* representing the *drift*, B is a linear operator, and dW is the differential of a Wiener process. We highlight the following facts. The state space of $X(t)$ is an infinite-dimensional space; the process W is as well an infinite-dimensional object, i.e. a vector-valued process; the drift f is a deterministic driving force of the dynamics; and $B\,dW$ shapes a stochastic driving force, which is a well-defined object shown with a vector-valued stochastic integration theory.

The solution of (1.2) is given by a stochastic version of the variation of parameters formula,

$$X(t) = S(t)u + \int_0^t S(t-s)f(s)\,ds + \int_0^t S(t-s)B\,dW(s), \tag{1.3}$$

where $S(t)$ is the semigroup generated by the differential operator $\mathcal{L}$.

It is noteworthy that in (1.3), the stochastic noise is still Gaussian (from the theorem that integration of a deterministic process with respect to a Wiener process is a Gaussian process), but no longer 'white'. The parameters of interest in the model (1.2) are the initial condition u and the drift f. The structures of these two parameters have different levels of similarity to the noise pattern. Hence, depending on the parameter to estimate, different approaches are more preferable.

The inference for evolution equations is the main subject of Part III.

1.3 Bayesian Methodology

In short, Bayesian methods utilize Bayes' rule to achieve the goal of inference. The term *Bayesian nonparametrics* refers to the Bayesian methods for infinite-dimensional models. In the Bayesian framework, a *prior distribution* (a probability measure) Π is assigned to the parameter θ. As a consequence, the prior induces a probability measure on the statistical model $\{\mathbb{P}_\theta : \theta \in \Theta\}$, where $\mathbb{P}_\theta$ is the law of X given θ. The *posterior distribution* is the conditional distribution of θ given X. The Bayesian procedure is identical for both parametric and nonparametric cases. However, the subtlety of Bayesian nonparametrics is higher, because of measurability concerns, and it will be considered in a subsequent chapter.

Our perspective on the Bayesian framework is that it offers a universal approach for nonparametric inference, while there are several issues necessary to address in order to obtain reasonable asymptotic results. Since we treat the Bayesian framework as a methodological device, and we are interested in the asymptotics of Posterior distributions, our standing is still in the frequentist regime. In other words, we want to understand the asymptotic performance of Bayesian methods from a frequentist perspective. In this thesis we solely focus on the goal just mentioned, and we have no intention to go further towards the long lasting disputation about the two regimes of frequentist and Bayesian at the philosophical level.

A more detailed review on Bayesian nonparametrics is given in Chapter 4.

1.4 Overview

This thesis is organized as follows.

In Part I, we prepare the basic elements that will serve as the building blocks for the later study on Bayesian inference for Gaussian linear models. Chapter 2 deals with the smoothness classes that will be used as the parameter spaces for the statistical study in the later chapters. Chapter 3 treats Gaussian measures on Banach spaces, which is the noise structure in all the models considered in this thesis. Chapter 4 presents the Bayesian nonparametric framework. We explore the noise structure of Gaussian linear models, with continuous and discrete observations, using the results from Chapter 3. In particular, we develop a posterior contraction theorem suitable for Gaussian linear models.

In Part II, we study Bayesian inference for linear inverse problems with Gaussian noise. We consider two types of problems, distinguished by their observation schemes: continuous and discrete observations, corresponding to the white noise model and regression model with transformed signals. In Chapter 5, we formally formulate the smoothing property of the forward operator $\mathcal{A}$, or equivalently, the *ill-posedness* of its inverse $\mathcal{A}^\dagger$, in the framework introduced in Chapter 2. In

Chapter 6, we systematically investigate the inverse problem with continuous observations. For the inverse problem with discrete observations, it is studied with two approaches. First, in Chapter 7, the inverse problem is studied in a concrete setting, by levering the Gaussian conjugacy in linear models. Then, in Chapter 8, we generalise the methodology developed in Chapter 6 to study the regression model.

In Part III, we study the inference for evolution equations. In Chapter 9, we present the semigroup approach to SPDEs, and as well introduce the additional structure suitable for statistical study. Subsequently, in Chapter 10 the Bayesian approach for the recovery of the parameters of evolution equations is examined.

1.5 Notations

We outline our conventions on the notations used in this work.

- The sets of Natural numbers, real numbers and complex numbers are $\mathbb{N} = \{1, 2, 3, \cdots\}$,$\mathbb{R}$ and $\mathbb{C}$, respectively. The *imaginary* unit is denoted by i. Special subsets are $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$, $\mathbb{R}^+ = (0, \infty)$ and $\mathbb{R}^+_0 = [0, \infty)$. Other similar notations are defined accordingly.

- $\mathbb{R}^d$ vectors and $\mathbb{N}^d$ multi-indices are denoted by $k = (k_1, \cdots, k_d)$. When a multi-index has identical entries, the following convention is adopted: $\boldsymbol{\beta} = (\beta, \cdots, \beta) \in \mathbb{R}^d_+$. For $p \in (0, \infty]$, the canonical p-norm (quasi-norm when $p < 1$) is defined as $|k|_p := (\sum_{i=1}^d k_i^p)^{1/p}$ with the usual modification for the case $p = \infty$. When $p = 2$, the norm is the standard Euclidean norm on $\mathbb{R}^d$. In this situation, the subscript is often omitted and we simply use $|\cdot|$. The following convention[5] is also used $k^\beta = (k^{\beta_1}, \cdots, k^{\beta_d})$. Partial orders are denoted by $\leq, \geq, \cdots$. For example, $j \leq k$ is understood as $j_i \leq k_i$, for all $i = 1, \cdots, d$.

- Constants are usually designated by capital letters I, J, M, N, etc. Index sets, domains in $\mathbb{R}^d$, are denoted by Fraktur letters $\mathfrak{I}, \mathfrak{J}, \mathfrak{D}$, etc.

- Real vector spaces are normally denoted by capital letters H, G, X, Y, etc., and their elements by small letters h, g, x, y, etc. Less often, blackboard bold is also used to designate spaces in the following situations. First, the elements in the space are customarily denoted by capital letters. One example is the sample space $\mathbb{X}$, which contains the observations customarily denoted by X. The other situation is when the letter indicates a particular space, for example, the reproducing kernel Hilbert space $\mathbb{H}$ of a Gaussian measure.

- The symbols $\lesssim, \gtrsim, \simeq$ mean $\leq, \geq, =$ up to a positive multiple independent of n (or another asymptotic parameter). The constant may be stated explicitly in subscripts, and e.g. $\lesssim_f$ means that it depends on f.

- For a normed space $(E, \|\cdot\|)$, the closed unit ball is denoted by $U(E) = \{x \in E : \|x\| \leq 1\}$. Given the topological dual E^* of E, the duality pair

---

[5] Notice its difference with the notation $k^\beta = \prod_{i=1}^d k_i^{\beta_i}$ sometimes appeared in literature.

is denoted by by $\langle \cdot, \cdot \rangle : E^* \times E \to \mathbb{R}$. The same notation is as well used for inner products on Hilbert spaces. The notation is consistent, since for real spaces sesquilinear form reduces to bilinear form. If there is a danger of confusion, subscripts are used.

Consider function spaces consisting of the functions with domain $\mathfrak{D}$ and codomain a vector space H. The function spaces are denoted similarly as for the real-valued case, e.g $L^2(\mathfrak{D}; H)$ the space of square integrable H-valued functions, i.e.

$$\int_{\mathfrak{D}} \|f\|_H^2 \, dx < \infty.$$

For $\mathbb{R}^d$-valued function spaces, the codomain is often omitted.

- General operators are designated by capital calligraphic letters $\mathcal{A}, \mathcal{T}, \cdots$ and Greek letters $\Lambda, \Phi, \cdots$, while there are exceptional cases, e.g. the expectation operator $\mathbb{E}$, identity mapping id, embedding ι, etc.
- Let μ and ν be two measures. If μ is dominated by (i.e. absolutely continuous to) ν, then it is denoted as $\mu \ll \nu$. If $\mu \ll \nu$ and $\nu \ll \mu$, i.e. they are equivalent measures, then we write $\mu \sim \nu$. Mutual singularity is denoted by $\mu \perp\!\!\!\perp \nu$.
- The following abbreviations are used in this thesis.

almost everywhere	a.e.
almost sure	a.s.

1.6 Notes

Asymptotic statistics, function spaces, are well established research fields and there exist numerous outstanding references. We only list a very small collection here, which by no means intends to be complete or exclusive. It is merely based on the author's familiarity.

Asymptotic Statistics
While written in 1940s, Cramér's book [22] still in large captures the essence of asymptotic theory, and additionally it well reflects the development of asymptotic theory at the early stage. Le Cam's noted treatise [66] largely presents the whole picture of asymptotic methods up to 1980s. Van der Vaart's textbook [97] is another standard reference in asymptotic theory, which provides a comprehensive introduction on the subject and also includes the new developments in 1990s. The more recent monograph [36] systematically depicts the asymptotic theory in nonparametric statistics. For the more detailed literature review, we refer to [66] for results up to 1980s and [36] for nonparametric asymptotics.

Function Spaces

Functions and function spaces serve as the fundamental element for many mathematical studies, and its own study has become a independent field long time ago. In this thesis, we do not use any advanced function spaces. Many classical results of the modern theory of function spaces can be found in [90]. The more recent contributions are largely collected in the sequel [91, 92, 93, 94] from the same author.

Separability

One fact which has been less addressed is the separability of parameter space Θ. A topological space is called *separable* if it contains a countable, dense subset. The previous examples of function spaces $C[0,1]$ and $L^2[0,1]$ are both separable. To begin with, separability is important in terms of approximation. Since separability imposes the existence of a countable dense subset, many approximation results can be stated and proved using induction, without invoking axiom of choice (or equivalently, Zorn's lemma). In other words, the induction arguments can be translated into implementable numerical algorithms. If the underlying space is nonseparable, it is no longer necessarily true. Additionally, since separability is one of the fundamental assumptions for the development of probability theory in Banach spaces (see [67]), it is appropriate to adopt the same notion.

For a separable Hilbert space, there always exists a countable orthonormal basis. It serves as the a cornerstone for the development of L^2 estimation theory. In general, for an arbitrary separable Hilbert space H, an isometric isomorphism can be established between H and ℓ^2 using projection (see Sobolev smoothness for example). The separability in Banach spaces is more involved. A separable Banach space does not necessarily have a (Schauder) basis, while a Banach space with a Schauder basis is necessarily separable. More importantly, Banach spaces with Schauder bases also possess *approximation property*, which roughly means that elements in the space can be approximated by finite-dimensional subjects.

Bayesian Methodology

The monograph [35] published in 2017 provides an extensive and thoroughly survey, covering almost all aspects, on the development of Bayesian nonparametrics up to the publication date. In addition, very comprehensive literature reviews are given in many places in the book. Therefore, here we simply refer to it for the reference on Bayesian nonparametrics.