Foundations of Nonparametric Bayesian Methods

Peter Orbanz

Overview: Today

1. Basic measure theory
2. Bayesian estimation
3. Construction of stochastic processes

Introduction: Parametric vs Nonparametric

Parametric model

A parameterized family of distributions, such that the number of parameters does not depend on sample size.

Nonparametric model

A parameterized model, but the number of parameters may grow with sample size.

Remarks:

- Number of parameters $\approx$ model complexity
- Complexity constant wrt sample size $\rightarrow$ nice convergence
- Typically: Nonparametric model $\rightarrow$ ∞-dim parameter space

Motivation: Measure Theory

Bayesian Nonparametrics

Probability models on infinite-dimensional spaces.

Problem: Density modeling

- Many ∞-dim distributions: No useful density.
- Some ∞-dim Bayesian models: No Bayes equation.

Measure-theoretic probability

- Most general available formalism for probability
- Measures good for proofs, densities good for modeling
- ∞-dim case: *Have to* work with measures

Measure Theory

Measure: Intuition

Roughly: Measure = Integral as a function of its region

$$\mu(A) = \int_A dx \qquad \text{or} \qquad \mu(A) = \int_A p(x)dx$$

Interpretation

$\mu(A)$ is mass of A, eg:

- Geometric case: Volume of A, or physical mass of a body.
- Probability case: Probability mass of event "random variable X takes value in A"

Integration: Abstract properties

Integrals: Decomposition properties

Write $\mu(A)$ for integral $\int_A dx$.

- $\mu(\emptyset) = 0$ (integral over empty set is zero)
- Pairwise disjoint sets A_n:

$$\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2) \quad \text{and} \quad \mu\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mu(A_n)$$

- If $B \subset A$:

$$\mu(B) \leq \mu(A) \qquad \text{and} \qquad \mu(A \setminus B) = \mu(A) - \mu(B)$$

Henri Lebesgue's Approach

Call any set function an integral (a measure) if it decomposes like an integral.

σ-algebras (1)

Motivation

- Defining measure: Often difficult/impossible on $\mathcal{P}(\Omega)$
- Idea: Restrict μ to subset $\mathcal{A}$ (“measurable sets”) of $\mathcal{P}(\Omega)$
- Measurable sets = sets over which we can integrate

Intuition: σ-algebra

- Always assume we can integrate over Ω
- If integrals on $A_1, A_2, \ldots$ given: Write $\mathcal{A} = \sigma(\{A_1, A_2, \ldots\})$ for system of all sets with deducable integrals.
- Compeleted set system $\mathcal{A}$ is called *σ-algebra*.

σ-algebras (2)

Def: σ-algebra

A system of sets $\mathcal{A} \subset \mathcal{P}(\Omega)$ is called a σ-algebra if:

1. $\emptyset, \Omega \in \mathcal{A}$
2. If $A \in \mathcal{A}$, then $\complement A \in \mathcal{A}$
3. If $A_n \in \mathcal{A}$ (for $n \in \mathbb{N}$), then $\cup_n^\infty A_n \in \mathcal{A}$

Constructing σ-algebras

Most important method:

- Start with: $\mathcal{T}$ = all open sets in Ω.
- σ-algebra: $\mathcal{B}(\Omega) := \sigma(\mathcal{T})$
  Read: $\sigma(\mathcal{T})$ = smallest σ-algebra that includes $\mathcal{T}$
- $\mathcal{B}(\Omega)$ is called the *Borel σ-algebra* of Ω
- Contains all open and closed sets

Measures

Def: Measure

Given σ-algebra $\mathcal{A}$, a *measure* is a function $\mu : \mathcal{A} \to \mathbb{R}_+$ with:

1. $\mu(\emptyset) = 0$
2. $\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$ if $A_n \in \mathcal{A}$ pairwise disjoint.

μ is *probability measure* if additionally

3. $\mu(\Omega) = 1$

Note: (1) and (2) imply all integral decomposition properties.

Most important measures

- *Lebesgue measure*: “Flat” measure on $\mathbb{R}^d$ (d-volume).
- *Counting measure*: $|A|$ if A finite set, $+\infty$ otherwise.

Densities

Intuition

Density = function that transforms measure μ_1 into
measure μ_2 by pointwise reweighting (on Ω)

Derivative Notation

$$d\mu_2(x) = f(x)d\mu_1(x) \qquad \text{or} \qquad \frac{d\mu_2}{d\mu_1}(x) = f(x)$$

Motivation: f = a function that is integrated to obtain μ_2
$\rightarrow$ "derivative" of μ_2

Immediate Question:

Is there always a density for μ_1, μ_2 given?

Radon-Nikodym Theorem

Absolute Continuity

- "Reweighting" by density

$$\mu_2(A) = \int_A d\mu_2(x) = \int_A f(x) d\mu_1(x)$$

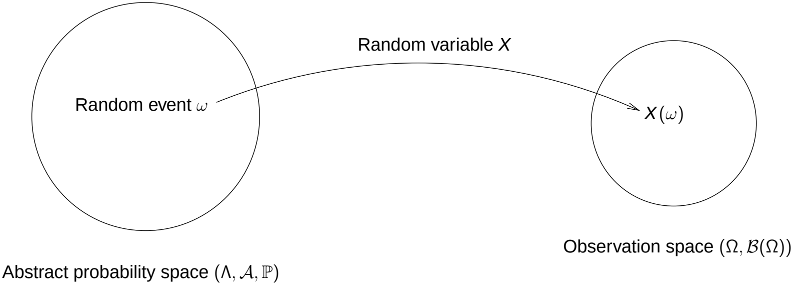
cannot work if $\mu_1(A) = 0$ and $\mu_2(A) \neq 0$.

- If that never happens for any $A \in \mathcal{A}$, then μ_2 is called "absolutely continuous wrt μ_1", in symbols: $\mu_2 \ll \mu_1$

Theorem (Radon-Nikodym)

Let μ_1, μ_2 be two finite measures on $\mathcal{A}$. Then μ_2 has a density w.r.t. μ_1 if and only if $\mu_2 \ll \mu_1$.

Probability: Formal Framework



- ω: atomic random event, “state of the universe”
- X: Random variable (mapping $\Lambda \to \Omega$)
- $X(\omega)$: observed random value
- $\mathbb{P}$: probability measure (distribution of ω)
- For set $A \in \mathcal{B}(\Omega)$: Probability of “$X(\omega) \in A$” = $\mathbb{P}(X^{-1}(A))$

Probability: Definitions

Def: Measurable mapping

Let $\mathcal{A}, \mathcal{B}$ be σ-algebras in Λ, Ω. A mapping $X : \Lambda \to \Omega$ is called *measurable* if $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$.

Interpretation: "F measurable" means that expression "$\mathbb{P}(X^{-1}(B))$" makes sense.

Def: Random variables

A *random variable* X is a measurable mapping from an abstract probability space $(\Lambda, \mathcal{A}, \mathbb{P})$ into an observation space $(\Omega, \mathcal{B}(\Omega))$.

Image Measure

The measure $\mathbb{P}$ is not known explicitly. We work with the distribution μ_X of random variable X defined as the *image measure*:

$$\mu_X := X(\mathbb{P}) \qquad \text{i.e.} \qquad \mu_X(A) := \mathbb{P}(X^{-1}(A))$$

Conditioning

Note

Defining conditional measures requires some effort.

Direct approach

Conditional probability of $X(\omega) \in A$ given that $X(\omega) \in B$:

$$\mu(A|B) := \frac{\mu(A \cap B)}{\mu(B)}$$

$\rightarrow$ no use if $\mu(B) = 0$ (think of Bayesian model on $\mathbb{R}^d$)

For now:

- We will just write $\mu(X|Y)$ for the conditional probability of X given Y and forget about details.
- If X, Y have a joint density, $\mu(X|Y)$ has a conditional density $p(x|y)$.

Parametric Model

Parametric model

Let $X : (\Lambda, \mathcal{A}) \to (\Omega_x, \mathcal{B}_x)$ and $\Theta : (\Lambda, \mathcal{A}) \to (\Omega_\theta, \mathcal{B}_\theta)$ be two random variables, and $\mu_X = X(\mathbb{P})$. Then the conditional distribution $\mu_X(X|\Theta)$ is called a *parametric family* of models (parameterized by $\theta \in \Omega_\theta$).

Bayesian model

If X observed and Θ unobserved, we call:

- $\mu_\Theta := \Theta(\mathbb{P})$ the *prior measure*
- $\mu_\Theta(\Theta|X)$ the *posterior measure*
- The overall model is called a Bayesian model.

Note: Not defined by a Bayes equation!

Bayes' Theorem

Problem:

Given the prior and the data, how can we determine the posterior? (Without exhaustive knowledge of $\mathbb{P}$, $\mathcal{A}$ etc)

Bayes Theorem

If the sampling model $\mu_X(X|\Theta)$ has density $p_{X|\theta}$, then:

$$\frac{d\mu_{\Theta|X}}{d\mu_{\Theta}}(\theta|x) = \frac{p_{X|\theta}}{\int p_{X|\theta} d\mu_{\theta}(\theta)}$$

for all x with $\int p_{X|\theta} d\mu_{\theta}(\theta) \notin \{0, \infty\}$.

Undominated Models

Dominated family:

Family of measures μ_t that all have density w.r.t. same ν.

In Bayes' theorem:

- "$p_{X|\theta}$ density of $\mu_{X|\Theta}$" requires family $\mu_{X|\Theta}$ dominated.
- Then: (1) posterior $\ll$ prior and (2) density generic.
- ∞-dim case: Often posterior $\ll$ prior *not* satisfied $\rightarrow$

  Bayesian model, but no Bayes equation.

Note:

"No Bayes equation" $\neq$ "intractable posterior"

Bayesian Nonparametrics

Nonparametric Bayesian model

A Bayesian model with:

1. $\dim(\Omega_\theta) = \dim(\Omega_x) = +\infty$.
2. Model can be evaluated on partial observations.

Partial observation

Random quantity with d dimensions, only $m < d$ are observed.

Example: GP regression

GP draw is function f, but only finite number of values of f known.

Stochastic Process Models

Intuition

Stochastic process = ∞-dim probability distribution

Typical GP definition

"A Gaussian process is a probability distribution on an infinite collection of random variables X_t such that the marginal distribution for each finite subset $(t_1, \ldots, t_n)$ of indices is Gaussian."

$\rightarrow$ Existence? Uniqueness?

Stochastic Process Construction (1)

Stochastic process measure μ^{E}: Distribution of RV

$$X^{\mathrm{E}} : (\Lambda, \mathcal{A}) \to (\Omega^{\mathrm{E}}, \mathcal{B}^{\mathrm{E}})$$

- E: infinite index set (indexes entries of random vector)
- Ω_0: "one-dimensional" sample space
- $\Omega^{\mathrm{E}} := \prod_{i \in E} \Omega_0$
- Interpretation: μ^{E}-draws = mappings $x : E \to \Omega_0$

Projector

P_{JI}:= projection mapping $\Omega^{\mathrm{J}} \to \Omega^{\mathrm{I}}$ (for $I \subset J \subset E$)

Marginals

Marginal of μ^{J} on $\Omega^{\mathrm{I}} \subset \Omega^{\mathrm{J}}$:

$$\underbrace{(\mathrm{P}_{\mathrm{JI}}\mu^{\mathrm{J}})(A)}_{\text{on } \Omega^{\mathrm{I}}} := \underbrace{\mu^{\mathrm{J}}(\mathrm{P}_{\mathrm{JI}}^{-1}A)}_{\text{on } \Omega^{\mathrm{J}}}$$

marginals = projections of measures

Stochastic Process Construction (2)

Def: Projective family

Family $\{\mu^I | I \subset E \text{ finite}\}$ such that for all finite I, J with $I \subset J$:

$$\mathrm{P}_{JI}\mu^J = \mu^I$$

Note: If μ^E given, the finite-dim marginals $\mu^I := \mathrm{P}_{EI}\mu^E$ are a projective family.

Kolmogorov's Extension Theorem

If a family $\{\mu^I | I \subset E \text{ finite}\}$ of finite-dimensional measures is projective, there exists a unique measure μ^E on Ω^E with μ^I as its marginals.

Jargon: μ^E is called the *projective limit* of the μ^I.

Example: GP construction

Choice of components

- $\Omega_0 := \mathbb{R}$ and index set $E = \mathbb{R}$
- P_{JI}: Euclidean projector from $\mathbb{R}^{|J|}$ to $\mathbb{R}^{|I|}$.
- Marginal family: μ^{I} are $|I|$-dimensional Gaussians

Ensure marginals projective

- Start with mean function $m(\,.\,)$ and covariance $k(\,.\,,\,.\,)$.
- Note: $E = \mathbb{R}$, finite $I = \{t_1, \dots, t_{|I|}\} \subset \mathbb{R}$
- μ^{I} = Gaussian, mean $(m(t_1), \dots, m(t_{|I|}))$ and $\Sigma_{ij} = k(t_i, t_j)$

Apply Extension Theorem

GP measure μ^{E} exists and is unique.

Note: μ^{E} has mean m and covariance function k, but that is *not* an immediate consequence of theorem!

The Problem with Kolmogorov

Problem

If dimension E is uncountable, the projective limit measure μ^{E} is basically useless.

Explanation

- Domain of μ^{E}: $\mathcal{B}^{\mathsf{E}}$ (generated by product topology)
- Sets in $\mathcal{B}^{\mathsf{E}}$: "axes-parallel" in all but countably many dimensions
- E uncountable $\rightarrow \mathcal{B}^{\mathsf{E}}$ too coarse for meaningful modeling

A Note of Caution:

Problem is often neglected in literature.
Example: Original paper on the DP (Ferguson, 1973).

Uncountable Dimensions

Intuition:

Objects of interest *effectively* have countably many degrees of freedom.

Examples

- Continuous functions: Completely defined by values on dense subset (e.g. $\mathbb{Q}$ in $\mathbb{R}$)
- Probability measures: Completely defined by values on countable system of sets.

Strategies

1. Modify theorem to directly define measure on “interesting” space (eg space of continuous functions).
2. Use Kolmogorov theorem, than restrict μ^{E} to interesting subspace.

Summary: Stochastic Process Construction

Kolmogorov

- Measure μ^{E} on product space Ω^{E} and "cylinder" σ-algebra $\mathcal{B}^{\mathsf{E}}$
- Conditions to check: Projective family
- Many interesting sets: Not product spaces
  product space $\leftrightarrow$ pointwise properties
- E uncountable: $\mathcal{B}^{\mathsf{E}}$ too coarse

Second Step

- If E countable: Done.
- If E uncountable: Measure μ^{E} has to be restricted to subspace to be useful.

Second step for uncountable E can be difficult.