Off the Beaten Path Tutorial:
Stochastic Processes and Simulations – Volume 1

Vincent Granville, Ph.D.
Data Shaping Solutions, LLC

Anacortes, WA, February 2022

Note: External links (in blue) and internal references (in red) are clickable throughout this document. Keywords highlighted in orange are indexed; those in red are both indexed and in the glossary section.

Contents

About this Textbook

This scratch course on stochastic processes covers significantly more material than usually found in traditional books or classes. The approach is original: I introduce a new yet intuitive type of random structure called perturbed lattice or Poisson-binomial process, as the gateway to all the stochastic processes. Such models have started to gain considerable momentum recently, especially in sensor data, cellular networks, chemistry, physics and engineering applications. I present state-of-the-art material in simple words, in a compact style, including new research developments and open problems. I focus on the methodology and principles, providing the reader with solid foundations and numerous resources: theory, applications, illustrations, statistical inference, references, glossary, educational spreadsheet, source code, stochastic simulations, original exercises, videos and more.

Below is a short selection highlighting some of the topics featured in the textbook. Some are research results published here for the first time.

GPU clustering	Fractal supervised clustering in GPU (graphics processing unit) using image filtering techniques aking to neural networks, automated black-box detection of the number of clusters, unsupervised clustering in GPU using density (gray levels) equalizer
Inference	New test of independence, spatial processes, model fitting, new type of confidence regions, minimum contrast estimation, oscillating estimators, mixture and superimposed models, radial cluster processes, exponential-binomial distribution with infinitely many parameters, generalized logistic distribution
Nearest neighbors	Statistical distribution of distances and Rayleigh test, Weibull distribution, properties of nearest neighbor graphs, size distribution of connected components, geometric features, hexagonal lattices, coverage problems, simulations, model-free inference
Cool stuff	Random functions, random graphs, random permutations, chaotic convergence, perturbed Riemann Hypothesis (experimental number theory), attractor distributions in extreme value theory, central limit theorem for stochastic processes, numerical stability, optimum color palettes, cluster processes on the sphere
Resources	27 Exercises with solution expanding the theory and methods presented in the textbook, well documented source code and formulas to generate various deviates and simulations, simple recipes (with source code) to design your own data animations as MP4 videos – see ours on YouTube

This first volume deals with point processes in one and two dimensions, including spatial processes and clustering. The next volume in this series will cover other types of stochastic processes, such as Brownian-related and random, chaotic dynamical systems. The point process which is at the core of this textbook is called the Poisson-binomial process (not to be confused with a binomial nor a Poisson process) for reasons that will soon become apparent to the reader. Two extreme cases are the standard Poisson process, and fixed (non-random) points on a lattice. Everything in between is the most exciting part.

Target Audience

College-educated professionals with an analytical background (physics, economics, finance, machine learning, statistics, computer science, quant, mathematics, operations research, engineering, business intelligence), students enrolled in a quantitative curriculum, decision makers or managers working with data scientists, graduate students, researchers and college professors, will benefit the most from this textbook. The textbook is also intended to professionals interested in automated machine learning and artificial intelligence.

It includes many original exercises requiring out-of-the-box thinking, and offered with solution. Both students and college professors will find them very valuable. Most of these exercises are an extension of the core material. Also, a large number of internal and external references are immediately accessible with one click, throughout the textbook: they are highlighted respectively in red and blue in the text. The material is organized to facilitate the reading in random order as much as possible and to make navigation easy. It is written for busy readers.

The textbook includes full source code, in particular for simulations, image processing, and video generation. You don't need to be a programmer to understand the code. It is well documented and easy to read, even for people with little or no programming experience. Emphasis is on good coding practices. The goal is to help you quickly develop and implement your own machine learning applications from scratch, or use the ones offered in the textbook. The material also features professional-looking spreadsheets allowing you to perform interactive statistical tests and simulations in Excel alone, without statistical tables or any coding. The code, data sets, videos and spreadsheets are available on my GitHub repository.

About the Author

Vincent Granville, PhD is a pioneering data scientist and machine learning expert, co-founder of Data Science Central (acquired by a publicly traded company in 2020), former VC-funded executive, author and patent owner. Vincent's past corporate experience includes Visa, Wells Fargo, eBay, NBC, Microsoft, CNET, InfoSpace and other Internet startup companies (one acquired by Google). Vincent is also a former post-doct from Cambridge University, and the National Institute of Statistical Sciences (NISS). He is currently publisher at DataShaping.com. He makes a living as an independent researcher working on stochastic processes, dynamical systems, experimental math and probabilistic number theory.

Vincent published in Journal of Number Theory, Journal of the Royal Statistical Society (Series B), and IEEE Transactions on Pattern Analysis and Machine Intelligence, among others. He is also the author of multiple books, including "Statistics: New Foundations, Toolbox, and Machine Learning Recipes", "Applied Stochastic Processes, Chaos Modeling, and Probabilistic Properties of Numeration Systems" with a combined reach of over 250,000, as well as "Becoming a Data Scientist" published by Wiley. For details, see my Google Scholar profile, here.

1 Poisson-binomial or Perturbed Lattice Process

I introduce here one of the simplest point process models. The purpose is to illustrate, in simple English, the theory of point processes using one of the most elementary and intuitive examples, keeping applications in mind. Many other point processes will be covered in the next sections, both in one and two dimensions. Key concepts, soon to be defined, include:

Category	Description	Book sections
Top parameters	Intensity λ – granularity of the process	4.4, 3.2.1
	Scaling factor s – quantifies point repulsion or mixing	3.1.1, 3.2.1
	Distribution F – location-scale family, with $F_s(x) = F(x/s)$	1.1, 3.2.2
Properties	Stationarity and ergodicity	1.4, 5.3
	Homogeneity and anisotropy	1.4.4
	Independent increments	1.4.3, 3.1.3
Core distributions	Interarrival times T	1.2, 4.2
	Nearest neighbor distances	3.4, 5.4
	Point count $N(B)$ in a set B	4.3, 5.3
	Point distribution (scattering, on a set B)	1.2
Type of process	Marked point process	1.5.1
	Cluster point process	2.1, 2.1.2
	Mixtures and interlacings (superimposed processes)	1.5.3, 3.4.3
Topology	Lattice space – index space divided by λ	2.1, 4.7
	State space – where the points are located	2.1
	Index space – index attached to a point (hidden process)	4.7, 2.2
Other concepts	Convergence to stationary Poisson point process	1.3, 4.6
	Boundary effects	3.5
	Dimension (of state space)	1.2
	Model identifiability	3.3

I also present several probability distributions that are easy to sample from, including logistic, uniform, Laplace and Cauchy. I use them in the simulations. I also introduce new ones such as the exponential-binomial distribution (the distribution of interarrival times), and a new type of generalized logistic distribution. One of the core distributions is the Poisson-binomial with an infinite number of parameters. The Poisson-binomial process is named after that distribution, attached to the point count (a random variable) counting the number of points found in any given set. By analogy, the Poisson point process is named after the Poisson distribution for its point count. Poisson-binomial processes are also known as perturbed lattice point processes. Lattices, also called grids, are a core topic in this textbook, as well as nearest neighbors.