An invitation to Bayesian nonparametrics

Nils Lid Hjort, Chris Holmes, Peter Müller and Stephen G. Walker

This introduction explains why you are right to be curious about Bayesian nonparametrics – why you may actually need it and how you can manage to understand it and use it. We also give an overview of the aims and contents of this book and how it came into existence, delve briefly into the history of the still relatively young field of Bayesian nonparametrics, and offer some concluding remarks about challenges and likely future developments in the area.

Bayesian nonparametrics

As modern statistics has developed in recent decades various dichotomies, where pairs of approaches are somehow contrasted, have become less sharp than they appeared to be in the past. That some border lines appear more blurred than a generation or two ago is also evident for the contrasting pairs "parametric versus nonparametric" and "frequentist versus Bayes." It appears to follow that "Bayesian nonparametrics" cannot be a very well-defined body of methods.

What is it all about?

It is nevertheless an interesting exercise to delineate the regions of statistical methodology and practice implied by constructing a two-by-two table of sorts, via the two "factors" parametric–nonparametric and frequentist–Bayes; Bayesian nonparametrics would then be whatever is not found inside the other three categories.

(i) *Frequentist parametrics* encompasses the core of classical statistics, involving methods associated primarily with maximum likelihood, developed in the 1920s and onwards. Such methods relate to various optimum tests, with calculation of p-values, optimal estimators, confidence intervals, multiple comparisons, and so forth. Some of the procedures stem from exact probability calculations for models that are sufficiently amenable to mathematical derivations, while others relate

to the application of large-sample techniques (central limit theorems, delta methods, higher-order corrections involving expansions or saddlepoint approximations, etc.).

(ii) *Bayesian parametrics* correspondingly comprises classic methodology for prior and posterior distributions in models with a finite (and often low) number of parameters. Such methods, starting from the premise that uncertainty about model parameters may somehow be represented in terms of probability distributions, have arguably been in existence for more than a hundred years (since the basic theorem that drives the machinery simply says that the posterior density is proportional to the product of the prior density with the likelihood function, which again relates to the Bayes theorem of *c*. 1763), but they were naturally limited to a short list of sufficiently simple statistical models and priors. The applicability of Bayesian parametrics widened significantly with the advent and availability of modern computers, from about 1975, and then with the development of further numerical methods and software packages pertaining to numerical integration and Markov chain Monte Carlo (MCMC) simulations, from about 1990.

As for category (i) above, asymptotics is often useful for Bayesian parametrics, partly for giving practical and simple-to-use approximations to the exact posterior distributions and partly for proving results of interest about the performance of the methods, including aspects of similarity between methods arising from frequentist and Bayesian perspectives. Specifically, frequentists and Bayesians agree in most matters, to the first order of approximation, for inference from parametric models, as the sample size increases. The mathematical theorems that in various ways make such statements precise are sometimes collectively referred to as "Bernshteĭn–von Mises theorems"; see, for example, Le Cam and Yang (1990, Chapter 7) for a brief treatment of this theme, including historical references going back not only to Bernshteĭn (1917) and von Mises (1931) but all the way back to Laplace (1810). One such statement is that confidence intervals computed by the frequentists and the Bayesians (who frequently call them "credibility intervals"), with the same level of confidence (or credibility), become equal, to the first order of approximation, with probability tending to one as the sample size increases.

(iii) *Frequentist nonparametrics* is a somewhat mixed bag, covering various areas of statistics. The term has historically been associated with test procedures that are or asymptotically become "distribution free," leading also to nonparametric confidence intervals and bands, etc.; for methodology related to statistics based on ranks (see Lehmann, 1975); then progressively with estimation of probability densities, regression functions, link functions etc., without parametric assumptions; and also with specific computational techniques such as the bootstrap. Again, asymptotics plays an important role, both for developing fruitful approximations

and for understanding and comparing properties of performance. A good reference book for learning about several classes of these methods is Wasserman (2006).

(iv) What ostensibly remains for our fourth category, then, that of *Bayesian nonparametrics*, are models and methods characterized by (a) big parameter spaces (unknown density and regression functions, link and response functions, etc.) and (b) construction of probability measures over these spaces. Typical examples include Bayesian setups for density estimation (in any dimension), nonparametric regression with a fixed error distribution, hazard rate and survival function estimation for survival analysis, without or with covariates, etc. The divisions between "small" and "moderate" and "big" for parameter spaces are not meant to be very sharp, and the scale is interpreted flexibly (see for example Green and Richardson, 2001, for some discussion of this).

It is clear that category (iv), which is the focus of our book, must meet challenges of a greater order than do the other three categories. The mathematical complexities are more demanding, since placing well-defined probability distributions on potentially infinite-dimensional spaces is inherently harder than for Euclidean spaces. Added to this is the challenge of "understanding the prior"; the ill-defined transformation from so-called "prior knowledge" to "prior distribution" is hard enough to elicit in lower dimensions and of course becomes even more challenging in bigger spaces. Furthermore, the resulting algorithms, for example for simulating unknown curves or surfaces from complicated posterior distributions, tend to be more difficult to set up and to test properly.

Finally, in this short list of important subtopics, we must note that the bigger world of nonparametric Bayes holds more surprises and occasionally exhibits more disturbing features than one encounters in the smaller and more comfortable world of parametric Bayes. It is a truth universally acknowledged that a statistician in possession of an infinity of data points must be in want of the truth – but some nonparametric Bayes constructions actually lead to inconsistent estimation procedures, where the truth is not properly uncovered when the data collection grows. Also, the Bernshteĭn–von Mises theorems alluded to above, which hold very generally for parametric Bayes problems, tend not to hold as easily and broadly in the infinite-dimensional cases. There are, for example, important problems where the nonparametric Bayes methods obey consistency (the posterior distribution properly accumulates its mass around the true model, with increased sample size), but with a different rate of convergence than that of the natural frequentist method for the same problem. Thus separate classes of situations typically need separate scrutiny, as opposed to theories and theorems that apply very grandly.

It seems clear to us that the potential list of good, worthwhile nonparametric Bayes procedures must be rather longer than the already enormously long lists of Bayes methods for parametric models, simply because bigger spaces contain more

than smaller ones. A book on Bayesian nonparametrics must therefore limit itself to only some of these worthwhile procedures. A similar comment applies to the *study* of these methods, in terms of performance, comparisons with results from other approaches, and so forth (making the distinction between the construction of a method and the study of its performance characteristics).

Who needs it?

Most modern statisticians have become well acquainted with various nonparametric and semiparametric tools, on the one hand (nonparametric regression, smoothing methods, classification and pattern recognition, proportional hazards regression, copulae models, etc.), and with the most important simulation tools, on the other (rejection–acceptance methods, MCMC strategies like the Gibbs sampler and the Metropolis algorithm, etc.), particularly in the realm of Bayesian applications, where the task of drawing simulated realizations from the posterior distribution is the main operational job. The *combination* of these methods is becoming increasingly popular and important (in a growing number of ways), and each such combination may be said to carry the stamp of Bayesian nonparametrics.

One reason why combining nonparametrics with Bayesian posterior simulations is becoming more important is related to practical feasibility, in terms of software packages and implementation of algorithms. The other reason is that such solutions contribute to the solving of actual problems, in a steadily increasing range of applications, as indicated in this book and as seen at workshops and conferences dealing with Bayesian nonparametrics. The steady influx of good real-world application areas contributes both to the sharpening of tools and to the sociological fact that, not only hard-core and classically oriented statisticians, but also various schools of other researchers in quantitative disciplines, lend their hands to work in variations of nonparametric Bayes methods. Bayesian nonparametrics is used by researchers working in finance, geosciences, botanics, biology, epidemiology, forestry, paleontology, computer science, machine learning, recommender systems, to name only some examples.

By prefacing various methods and statements with the word "Bayesian" we are already acknowledging that there are different schools of thought in statistics – Bayesians place prior distributions over their parameter spaces while parameters are fixed unknowns for the frequentists. We should also realize that there are different trends of thought regarding how statistical methods are actually used (as partly opposed to how they are constructed). In an engaging discussion paper, Breiman (2001) argues that contemporary statistics lives with a Snowean "two cultures" problem. In some applications the careful study and interpretation of finer aspects of the model matter and are of primary concern, as in various substantive

sciences – an ecologist or a climate researcher may place great emphasis on determining that a certain statistical coefficient parameter is positive, for example, as this might be tied to a scientifically relevant finding that a certain background factor really influences a phenomenon under study. In other applications such finer distinctions are largely irrelevant, as the primary goals of the methods are to make efficient predictions and classifications of a sufficient quality. This pragmatic goal, of making good enough "black boxes" without specific regard to the components of the box in question, is valid in many situations – one might be satisfied with a model that predicts climate parameters and the number of lynx in the forest, without always needing or aiming to understand the finer mechanisms involved in these phenomena.

This continuing debate is destined to play a role also for Bayesian nonparametrics, and the right answer to what is more appropriate, and to what is more important, will be largely context dependent. A statistician applying Bayesian nonparametrics may use one type of model for uncovering effects and another for making predictions or classifications, even when dealing with the same data. Using different models for different purposes, even with the very same data set, is not a contradiction in terms, and relates to different loss functions and to themes of interest-driven inference; cf. various focused information criteria for model selection (see Claeskens and Hjort, 2008, Chapter 6).

It is also empirically true that some statistics problems are easier to attack using Bayesian methods, with machineries available that make analysis and inference possible, in the partial absence of frequentist methods. This picture may of course shift with time, as better and more refined frequentist methods may be developed also, for example for complex hierarchical models, but the observation reminds us that there is a necessary element of pragmatism in modern statistics work; one uses what one has, rather than spending three extra months on developing alternative methods. An eclectic view of Bayesian methods, prevalent also among those statisticians hesitant to accept all of the underlying philosophy, is to use them nevertheless, as they are practical and have good performance. Indeed a broad research direction is concerned with reaching performance-related results about classes of nonparametric Bayesian methods, as partly distinct from the construction of the models and methods themselves (cf. Chapter 2 and its references). For some areas in statistics, then, including some surveyed in this book, there is an "advantage Bayes" situation. A useful reminder in this regard is the view expressed by Art Dempster: "a person cannot be Bayesian or frequentist; rather, a particular *analysis* can be Bayesian or frequentist" (see Wasserman, 2008). Another and perhaps humbling reminder is Good's (1959) lower bound for the number of different Bayesians (46 656, actually), a bound that may need to be revised upwards when the discussion concerns nonparametric Bayesians.

Why now?

Themes of Bayesian nonparametrics have engaged statisticians for about forty years, but now, that is around 2010, the time is ripe for further rich developments and applications of the field. This is due to a confluence of several different factors: the availability and convenience of computer programs and accessible software packages, downloaded to the laptops of modern scientists, along with methodology and machinery for finessing and finetuning these algorithms for new applications; the increasing accessibility of statistical models and associated methodological tools for taking on new problems (leading also to the development of further methods and algorithms); various developing application areas paralleling statistics that find use for these methods and sometimes develop them further; and the broadening meeting points for the two flowing rivers of nonparametrics (as such) and Bayesian methods (as such).

Evidence of the growing importance of Bayesian nonparametrics can also be traced in the archives of conferences and workshops devoted to such themes. In addition to having been on board in broader conferences over several decades, an identifiable subsequence of workshops and conferences set up for Bayesian nonparametrics per se has developed as follows, with a rapidly growing number of participants: Belgirate, Italy (1997), Reading, UK (1999), Ann Arbor, USA (2001), Rome, Italy (2004), Jeju, Korea (2006), Cambridge, UK (2007), Turin, Italy (2009). Monitoring the programs of these conferences one learns that development has been and remains steady, regarding both principles and practice.

Two more long-standing series of workshops are of interest to researchers and learners of nonparametric Bayesian statistics. The BISP series (Bayesian inference for stochastic processes) is focused on nonparametric Bayesian models related to stochastic processes. Its sequence up to the time of writing reads Madrid (1998), Varenna (2001), La Mange (2003), Varenna (2005), Valencia (2007), Brixen (2009), alternating between Spain and Italy. Another related research community is defined by the series of research meetings on objective Bayes methodology. The coordinates of the O'Bayes conference series history are Purdue, USA (1996), Valencia, Spain (1998), Ixtapa, Mexico (2000), Granada, Spain (2002), Aussois, France (2003), Branson, USA (2005), Rome, Italy (2007), Philadelphia, USA (2009).

The aims, purposes and contents of this book

This book has in a sense grown out of a certain event. It reflects this particular origin, but is very much meant to stand solidly and independently on its constructed feet, as a broad text on modern Bayesian nonparametrics and its theory and methods; in other words, readers do not need to know about or take into account the event that led to the book being written.

A background event

The event in question was a four-week program on Bayesian nonparametrics hosted by the Isaac Newton Institute of Mathematical Sciences at Cambridge, UK, in August 2007, and organized by the four volume editors. In addition to involving a core group of some twenty researchers from various countries, the program organized a one-week international conference with about a hundred participants. These represented an interesting modern spectrum of researchers whose work in different ways is related to Bayesian nonparametrics: those engaged in methodological statistics work, from university departments and elsewhere; statisticians involved in collaborations with researchers from substantive areas (like medicine and biostatistics, quantitative biology, mathematical geology, information sciences, paleontology); mathematicians; machine learning researchers; and computer scientists.

For the workshop, the organizers selected four experts to provide tutorial lectures representing four broad, identifiable themes pertaining to Bayesian nonparametrics. These were not merely four themes "of interest," but were closely associated with the core models, the core methods, and the core application areas of nonparametric Bayes. These tutorials were

- Dirichlet processes, related priors and posterior asymptotics (by S. Ghosal),
- models beyond the Dirichlet process (by A. Lijoi),
- applications to biostatistics (by D. B. Dunson),
- applications to machine learning (by Y. W. Teh).

The program and the workshop were evaluated (by the participants and other parties) as having been very successful, by having bound together different strands of work and by perhaps opening doors to promising future research. The experience made clear that nonparametric Bayes is an important growth area, but with side-streams that may risk evolving too much in isolation if they do not make connections with the core field. All of these considerations led to the idea of creating the present book.

What does this book do?

This book is structured around the four core themes represented by the tutorials described above, here appearing in the form of invited chapters. These core chapters are then complemented by chapters written by the four volume editors. The role of these complementary chapters is partly to discuss and extend the four core chapters, in suitably matched pairs. These complements also offer further developments and provide links to related areas. This editorial process hence led to the following list of chapters, where the pairs 1–2, 3–4, 5–6, 7–8 can be regarded as units.

1 S. G. Walker: Bayesian nonparametric methods: motivation and ideas
2 S. Ghosal: The Dirichlet process, related priors and posterior asymptotics
3 A. Lijoi and I. Prünster: Models beyond the Dirichlet process
4 N. L. Hjort: Further models and applications.
5 Y. W. Teh and M. I. Jordan: Hierarchical Bayesian nonparametric models with applications
6 J. Griffin and C. Holmes: Computational issues arising in Bayesian nonparametric hierarchical models
7 D. B. Dunson: Nonparametric Bayes applications to biostatistics
8 P. Müller and F. Quintana: More nonparametric Bayesian models for biostatistics

As explained at the end of the previous section, it would not be possible to have "everything important" inside a single book, in view of the size of the expanding topic. It is our hope and view, however, that the dimensions we have probed are sound, deep and relevant ones, and that different strands of readers will benefit from working their way through some or all of these.

The *first* core theme (Chapters 1 and 2) is partly concerned with some of the cornerstone classes of nonparametric priors, including the Dirichlet process and some of its relatives. General principles and ideas are introduced (in the setting of i.i.d. observations) in Chapter 1. Mathematical properties are further investigated, including characterizations of the posterior distribution, in Chapter 2. The theme also encompasses properties of the behavior of the implied posterior distributions, and, specifically, consistency and rates of convergence. Bayesian methodology is often presented as essentially a machinery for coming from the prior to the posterior distributions, but is at its most powerful when coupled with decision theory and loss functions. This is true in nonparametric situations as well, as also discussed inside this first theme.

The *second* main theme (Chapters 3 and 4) is mainly occupied with the development of the more useful nonparametric classes of priors beyond those related to the Dirichlet processes mentioned above. Chapter 3 treats completely random measures, neutral-to-the-right processes, the beta process, partition functions, clustering processes, and models for density estimation, with Chapter 4 providing further methodology for stationary time series with nonparametrically modeled covariance functions, models for random shapes, etc., along with pointers to various application areas, such as survival and event history analysis.

The third and fourth core themes are more application driven than the first two. The *third* core theme (Chapters 5 and 6) represents the important and growing area of both theory and applications of Bayesian nonparametric hierarchical modeling (an area related to what is often referred to as machine learning). Hierarchical

modeling, again with Dirichlet processes as building blocks, leads to algorithms that solve problems in information retrieval, multipopulation haplotype phasing, word segmentation, speaker diarization, and so-called topic modeling, as demonstrated in Chapter 5. The models that help to accomplish these tasks include Chinese restaurant franchises and Indian buffet processes, in addition to extensive use of Gaussian processes, priors on function classes such as splines, free-knot basis expansions, MARS and CART, etc. These constructions are associated with various challenging computational issues, as discussed in some detail in Chapter 6.

Finally the *fourth* main theme (Chapters 7 and 8) focuses on biostatistics. Topics discussed and developed in Chapter 7 include personalized medicine (a growing trend in modern biomedicine), hierarchical modeling with Dirichlet processes, clustering strategies and partition models, and functional data analysis. Chapter 8 elaborates on these themes, and in particular discusses random partition priors and certain useful variations on dependent Dirichlet processes.

How do alternative models relate to each other?

Some comments seem in order to put the many alternative models in perspective. Many of the models are closely related mathematically, with some being a special case of others. For example, the Dirichlet process is a special case of the normalized random measure with independent increments introduced in Chapter 3. Many of the models introduced in later chapters are natural generalizations and extensions of earlier defined models. Several of the models introduced in Chapter 5 extend the random partition models described in the first four chapters, including, for example, a natural hierarchical extension of the Dirichlet process model. Finally, Chapters 7 and 8 introduce many models that generalize the basic Dirichlet process model to one for multiple related random probability measures. As a guideline for choosing a model for a specific application, we suggest considering the data format, the focus of the inference, and the desired level of computational complexity.

If the data format naturally includes multiple subpopulations then it is natural to use a model that reflects this structure in multiple submodels. In many applications the inference of interest is on random partitions and clustering, rather than on a random probability measure. It is natural then to use a model that focuses on the random partitions, such as a species sampling model. Often the choice will simply be driven by the availability of public domain software. This favors the more popular models such as Dirichlet process models, Pólya tree models, and various dependent Dirichlet process models.

The reader may notice a focus on biomedical applications. In part this is a reflection of the history of nonparametric Bayesian data analysis. Many early papers

focused on models for event time data, leading naturally to biomedical applications. This focus is also a reflection of the research experience of the authors. There is no intention to give an exhaustive or even representative discussion of areas of application. An important result of focusing on models rather than applications is the lack of a separate chapter on hierarchical mixed effects models, although many of these feature in Chapters 7 and 8.

How to teach from this book

This book may be used as the basis for master's or Ph.D. level courses in Bayesian nonparametrics. Various options exist, for different audiences and for different levels of mathematical skill. One route, for perhaps a typical audience of statistics students, is to concentrate on core themes two (Chapters 3 and 4) and four (Chapters 7 and 8), supplemented with computer exercises (drawing on methods exhibited in these chapters, and using for example the software `DPpackage`, described in Jara, 2007). A course building upon the material in these chapters would focus on data analysis problems and typical data formats arising in biomedical research problems. Nonparametric Bayesian probability models would be introduced as and when needed to address the data analysis problems.

More mathematically advanced courses could include more of core theme one (Chapters 1 and 2). Such a course would naturally center more on a description of nonparametric Bayesian models and include applications as examples to illustrate the models. A third option is a course designed for an audience with an interest in machine learning, hierarchical modeling, and so forth. It would focus on core themes two (Chapters 2 and 3) and three (Chapters 5 and 6).

Natural prerequisites for such courses as briefly outlined here, and by association for working with this book, include a basic statistics course (regression methods associated with generalized linear models, density estimation, parametric Bayes), perhaps some survival analysis (hazard rate models, etc.), along with basic skills in simulation methods (MCMC strategies).

A brief history of Bayesian nonparametrics

Lindley (1972) noted in his review of general Bayesian methodology that Bayesians up to then had been "embarrassingly silent" in the area of nonparametric statistics. He pointed out that there were in principle no conceptual difficulties with combining "Bayesian" and "nonparametric" but indirectly acknowledged that the mathematical details in such constructions would have to be more complicated.