Research statement

B. J. K. Kleijn *Feb 2022*

Over the course of the past two decades the subject of non-parametric Bayesian statistics has seen rapid development, enjoying widening scope and rising popularity. Driven also by ever-increasing computational power and sophistication, applications and theory have increased steadily, now to the point where one can speak of a true field within statistics. A rough sub-division could be based on a (mostly purely Bayesian) sub-field focussed on Bayesian modelling with process priors, a (mostly computer-science oriented) sub-field of numerical/approximation methods for sampling of posterior distributions, and a (mostly frequentist) sub-field that studies the large-sample behaviour of non-parametric posterior distributions.

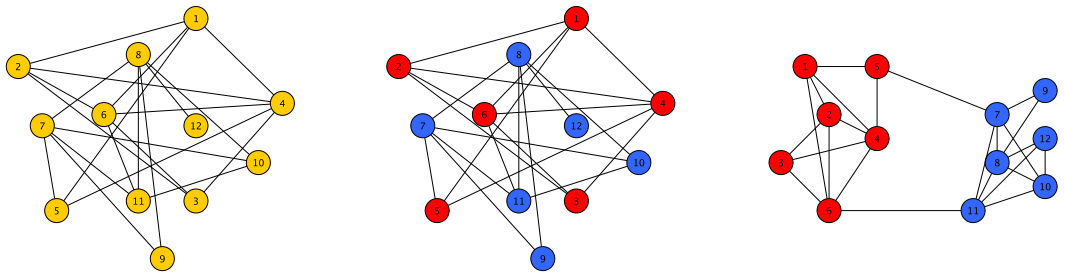
My work in those years has focussed on several subjects: *model misspecification* [TH2, 8, 9, 13] concerns the large-sample behaviour of posteriors for models that do not include the distribution that generated the data. With my thesis advisor Aad van der Vaart, minimax testing, posterior rates of convergence and the Bernstein-von Mises phenomenon were studied. At UC Berkeley (with a TALENT grant from NWO), I worked with Peter Bickel on a difficult subject, *Efficiency of semi-parametric Bayesian estimation methods* [12, 16, 18], aimed at the formulation of a Bernstein-von Mises theorem in non-parametric models where one is only interested in estimation of a functional, rather than the whole distribution of the data. This work was later supported by a VENI grant from NWO and continued with PhD-student Bartek Knapik [14]. Together with Yongdai Kim and PhD-student Minwoo Chae, another follow-up concerned the *Bernstein-von Mises phenomenon in regression models with symmetric errors* [18]. In a project with Aad van der Vaart and PhD-student Stéphanie van der Pas, a *Bayesian method for sparse variable selection* with the so-called *horseshoe prior* was studied [19].

Throughout, my personal interest has been focussed on *large-sample concentration of posterior distributions* [8, 9, 12, 13, 15, 16, 19, 23, 27]: as the amount of available data grows, one requires that posterior distributions become more informative by concentrating a larger-and-larger fraction of posterior probability in smaller-and-smaller neighbourhoods of the true distribution that generated the data. The centre piece of large-sample Bayesian statistics has always been Lorraine Schwartz's 1965 posterior consistency theorem, based on the existence of test sequences and the use of so-called Kullback-Leibler priors. My work on generalization of Schwartz's conditions started with a detailed study of *inconsistent posteriors* with Peter Bickel, Anthony Gamst and Ya'acov Ritov [15]. More direct was a variation on Schwartz's proof that merged the testing condition with the Kullback-Leibler condition, conducted in collaboration with Yanyun Zhao[†] [19]. Definitive answers are reached in [23], which combines a Bayesian variation on the test condition with a novel form of Le Cam's contiguity termed *remote contiguity* to replace the Kullback-Leibler condition. This results in new and simpler frequentist consistency theorems for posteriors, as well as consistent hypothesis-testing/model-selection with posterior odds and a general identification of (enlarged) credible sets as frequentist confidence sets. Like contiguity, the concept of remote contiguity is useful much more broadly and has already been used by other authors. Based on the above, I have compiled a book on *the frequentist theory of Bayesian statistics* [BK], which is in the final stages of writing and will be published

by Springer Verlag.

Besides these subjects in mathematical statistics, I have also worked on practical problems involving real-world data: together with Peter Bickel and John Rice, I have worked on the *detection of periodicity in astronomical point-processes*, to find signals of pulsars in high-energy photon data collected by NASA's satellite-borne EGRET gamma-ray detector. This has resulted in publications in astrophysics journals [10, 11] and an NSF AST grant. Over the past four years I have worked with PhD-student Mike Derksen (in collaboration with Robin de Vilder of hedge fund Deep Blue Capital NV) on *asset price formation and distributions for auction returns*, based on a newly developed stochastic model for price formation through supply-demand equilibria [20, 21], including a perspective on the origin of heavy tails in distributions of daily returns [22]. (It is worth mentioning that both [11] and [20] are considered A-journals in their respective disciplines.)

Motivated by great progress in machine learning, artificial intelligence, network science and data science, trends in statistics over the past years have been directed towards computational methods, emphasizing applications and methodology rather than theory. In mathematical statistics there has been a corresponding shift towards numerical approximation and a greater focus on increasingly detailed, model-specific calculations. For mathematical statisticians with expertise on the very mathematical side of the spectrum these developments have opened up possibilities on the intersections with probability theory and pure mathematics. To be more specific, I believe that there are very real and exciting opportunities to apply new ideas in probability theory and powerful methods from functional analysis to modern questions in (mathematical) statistics, in data and network science and in machine learning.



(a) Observed graph (b) Unobserved communities (c) Community detection

Figure 1: A realisation of the stochastic block graph (Fig. 1(a)) with $n = 12$ vertices from two unobserved communities: vertices 1 through 6 belong to the red community and vertices 6 through 12 to the blue (Fig. 1(b)). Community detection (Fig. 1(c)) estimates the communities of Fig. 1(b), based on the presence or absence of edges in Fig. 1(a).

My present and future research is aimed at some of the resulting interdisciplinary niches. To demonstrate with a concrete example, Jan van Waaij and I have considered the question of community detection in the two-community version of the so-called *stochastic block model* from network science: the observed data is a random graph in which edges are present randomly with probabilities that depend on the (unobserved) communities of the vertices that they connect; the statistical challenge is to recover the communities from observation of the graph (see Fig. 1). Recovery of communities in this model is considered one of the central questions in network science and many recovery algorithms have

been proposed. Probabilists have analysed the lower bounds on edge sparsity required to make recovery possible. With the methods of [23], van Waaij and I show that posteriors recover the community structure consistently even in the most sparse graphs [24, 25, 26, 27], and more importantly, we construct exact confidence sets for community structure from (enlarged) credible sets *for finite graph sizes* [27]. Finite-sample uncertainty quantification in modern problems from machine learning and network science has remained notoriously elusive so far, so the possibility to construct exact frequentist confidence sets from (computationally far more accessible) Bayesian credible sets is a uniquely important new tool from a methodological point of view. An NWO-ENW-M-1 proposal for a PhD project to generalize the construction and explore this possibility further has been submitted. Although NWO's final decision will not be made until halfway April 2022, the referee reports are unequivocally positive and supportive of the proposal.

Regarding the niche between pure mathematics and mathematical statistics, my efforts are essentially based on Le Cam's 1986 book. To read it, some basic understanding of the version of functional analysis of Nicolas Bourbaki, Laurent Schwartz and Francois Treves is required. Reading their works is a tremendous source of inspiration and has led to two large projects so far: first is the answer to the question, *which pairs of hypotheses are asymptotically testable and which are not?* [28]. A theorem of Lucien Le Cam and Lorraine Schwartz from 1960 enables the formulation of necessary and sufficient topological conditions for the existence of test sequences, without imposing conditions on the underlying model, for uniform, pointwise and Bayesian testability. The existence of uniform tests is closely related to the existence of adaptive confidence sets (a popular theme over the last decade); the existence of Bayesian tests forms one of the two basic conditions in [23] and determines whether posteriors concentrate; pointwise testability is a concept that both shapes the foundations of frequentist statistics, and forms the essence of many methodological questions (for example, model selection).

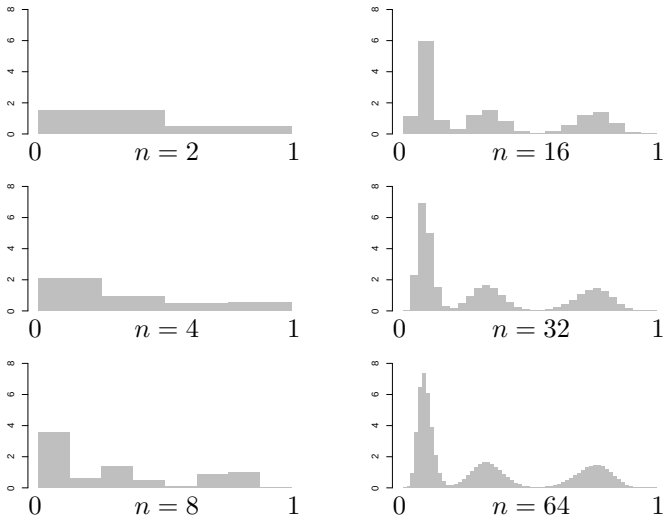


Figure 2: Refining histograms on the interval $[0, 1]$ for the mixture of Beta-distributions $\frac{1}{2}\text{Beta}(10, 100) + \frac{1}{4}\text{Beta}(20, 40) + \frac{1}{4}\text{Beta}(30, 10)$; if a system of distributions for random histograms is provided based on an infinite sequence of refinements, does this define a random probability distribution on $[0, 1]$?

Second is the formulation of existence theorems for random probability distributions defined as limits of refining systems of random histograms [29] (see Fig. 2). Examples are the well-known Dirichlet and Pólya tree systems, which are popular for their computational accessibility and their theoretical properties, not only in non-parametric Bayesian statistics, but more generally, in probability theory, network science and machine learning. The *existence question* for these so-called inverse limit probability measures is a difficult mathematical point that has retained the attention of (Bayesian) statisticians, probabilists and functional analysts ever since Ferguson's seminal work of the 1970's. Based on a theorem of Laurent Schwartz and Nicolas Bourbaki on inverse limit Radon measures, it is possible to give accessible conditions for existence, which reveal that there are three distinct phases for random histogram limits (called singular, Cantor and absolutely continuous respectively), depending on the degree of refinement of the model topology and directly relevant to the (computationally important) approximative properties of the histogram distributions. Dirichlet distributions are all in the Cantor phase, but the Pólya tree family has examples in all three phases. When applied to signed measures, said theorem proves the existence of what Wendelin Werner calls the Gaussian free field (a random tempered distribution, important in Euclidean quantum field theory). With an appeal to Martingale convergence, the theorem also proves existence of the ϕ^4 interacting bosonic field in four dimensions, answering a long-standing question in theoretical physics regarding so-called asymptotic triviality of said interacting field.

To summarize, my future work will be aimed at modern questions in statistics and probability through application of the ideas of [23], [28] and [29], with an emphasis on connections with stochastic analysis, functional analysis and topology, to broaden the fantastic perspective that stochastics offers and to explore the beauty of exact sciences in general.

(A 20-minute slide presentation of this research statement, with more detailed descriptions of associated publications, (PhD. and MSc.) students and (past, present and future) grant proposals is available on request, or can form the basis for an in-person presentation/discussion.)