SEPTEMBER 2025
VOL. 32 NO. 3

INTERNATIONAL SOCIETY FOR BAYESIAN ANALYSIS

THE ISBA BULLETIN

OFFICIAL BULLETIN OF THE INTERNATIONAL SOCIETY FOR BAYESIAN ANALYSIS

SOFTWARE HIGHLIGHT

Luca Porta Mana
pgl@portamana.org

INFERNO
INFERENCE IN R WITH BAYESIAN NONPARAMETRICS

1 Population inference and Bayesian nonparametrics

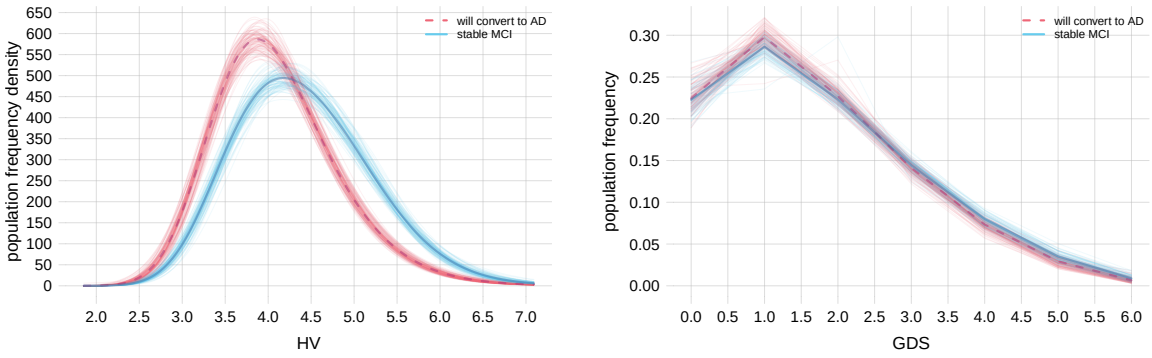
A very important kind of inference in research fields such as medicine is *population inference*, often also called "density inference" or "density regression". Its general goal is to infer the frequency distribution of some variates in a population. This is different, for instance, from *functional regression*, where the goal is to infer the functional relationship – assumed to exist – between a set of predictor variates and a target or "predictand" variate. In population inference the existence of a functional relation cannot be assumed. In fact there may not even be a clear distinction between predictor and predictand variates. A typical goal is the inference of frequency distributions within particular sub-populations or sub-groups, thus all kinds of conditional probabilities are required. A clinician may be interested in the statistics and probability of a medical condition given a symptom, but also that of a symptom given a medical condition; and maybe only within subjects of a given sex or age. De Finetti's theorem (see e.g. Bernardo and Smith, 2000, §§ 4.2, 4.3, 4.6) lies at the heart of population-inference methods; a particularly brilliant discussion is given by Lindley & Novick (1981).

Sadly many researchers still approach population-inference problems by means of p-values or other frequentist practices, which only give limited, coarse, and not seldom misleading results about a population's frequency distribution. Some researchers adopt Bayesian methods but limit themselves to *parametric* ones, which make very restrictive and possibly unrealistic assumptions about the population's distribution; as opposed to *nonparametric* ones, which don't.

Until a couple decades ago the use of parametric methods, and maybe even of frequentist practices, was somehow justified by pragmatic reasons. Better methods were computationally too costly or unfeasible. Population-inference problems were low-dimensional; one could *visually* check whether the assumptions were appropriate to the problem and the results reasonable. Today the reasons cannot be earnestly be given, however. Bayesian nonparametric methods have become computationally feasible for many kinds of population inference. And many population-inference problems today involve tens, hundreds, or thousands of variates of different kinds; being so high-dimensional, it is impossible to visually check whether frequentist practices or parametric assumptions are acceptable, or by how much they err. Results may therefore be affected by large errors (Draper, 1995), whose existence is often not even reported.

One reason can still be given today for the avoidance of Bayesian nonparametric methods for population inferences: *lack of user-friendly software*. Clinicians who'd be curious to try out a Bayesian nonparametric analysis of their studies simply can't: that would require the study of Markov-chain Monte Carlo techniques, of programming languages to implement the latter, and of a plethora of debated practices to "assess convergence". Most clinicians don't have time to learn all this even if they wanted to. Available packages for Bayesian nonparametrics are not quite suited to population inference. Some focus on functional regression, which as discussed above is not an appropriate assumption. Some make an a priori distinction between predictor and predictand variates, limiting the range of useful inferences. Most still require non-statistical technical expertise.

Figure 1



The R-package **inferno**[1] was built to try to remedy the lack of this kind of software.

2 Use and features

(In the following, for brevity, the term 'probability' will also stand for 'probability density' depending on nature of the quantities involved; similarly for 'frequency'.)

The application of the package is simple. The user provides a sample of variates D from a population, for instance the age, sex, symptom, disease, and kind of treatment of a number of patients. Thereafter, the user can calculate any of the following:

- For a new unit of the population, say a new patient, the conditional probability for the values of *any* set of variates $X{=}x, Y{=}y, \dots$ given *any* other set $Z{=}z, W{=}w, \dots$:

$$\mathrm{P}(X{=}x, Y{=}y, \dots \mid Z{=}z, W{=}w, \dots, D) \tag{1}$$

  Such a probability could for instance be used in medical decision making (Sox et al., 2024; Hunink et al., 2014). The conditional could also be empty.

- The probability distribution for the conditional *frequencies*, in the whole unsampled population, of the values of any set of variates given any other set. If we denote by F the frequency and by f a specific value,

$$\mathrm{P}\bigl[F(X{=}x, \dots \mid Z{=}z, \dots) = f \bigm| D\bigr] \,. \tag{2}$$

  This probability distribution is represented by Monte Carlo samples drawn from it.

- The mutual information (MacKay, 2005, Ch. 8) between any two sets of quantities.

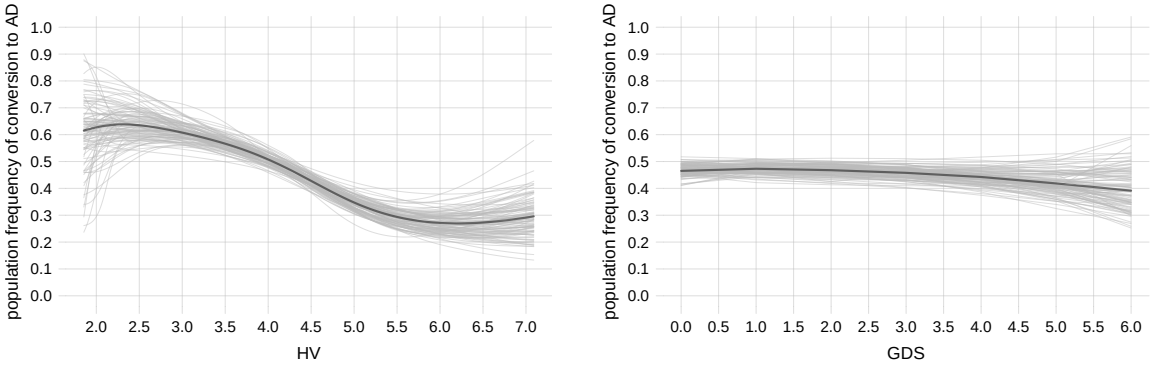
Probabilities (1) and (2) are connected by a variant of de Finetti's theorem:

$$\mathrm{P}(X{=}x, \dots \mid Z{=}z, \dots, D) = \int f \, \mathrm{P}\bigl[F(X{=}x, \dots \mid Z{=}z, \dots) = f \bigm| D\bigr] \, \mathrm{d}f \tag{3}$$

In a manner of speaking, the probability distribution (2) expresses how much the probability value (1) could change, if it were updated by sampling the whole population. It expresses the uncertainty in the statistical results owing to finite sample size.

[1] https://pglpm.github.io/inferno/

Figure 2



The package allows the user to visualize the probabilities above when just one "predictand" and one predictor variates are involved. Figure 1 shows two examples from a study (Mana et al., 2023) about conversion from Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD); they can be used to further illustrate the probability distributions discussed above.

In the plot on the left, the predictand variate is hippocampal volume (HV); the predictor is the binary, yes/no variate 'will convert to AD' (cAD). The solid blue thick line and dashed red thick line are the conditional probability distributions (omitting the data dependence)

$$\mathrm{P}(\mathrm{HV} \mid \mathrm{cAD}{=}\mathrm{N}) \,, \qquad \mathrm{P}(\mathrm{HV} \mid \mathrm{cAD}{=}\mathrm{Y}) \,.$$

The cloud of blue thinner lines that surrounds the first distribution above represents the probability distribution of possible frequency distributions

$$\mathrm{P}\big[F(\mathrm{HV} \mid \mathrm{cAD}{=}\mathrm{N})\big]$$

each thin line is a sample from such distribution.

Looking at this plot, a clinician can immediately see that the frequency distribution of hyppocampal volume is clearly different in the sub-populations of patients that will convert to Alzheimer's and those who won't. Such a difference is almost certain even considering the uncertainty from the finite sample size.

The plot on the right is analogous but for the predictand variate 'geriatric depression scale' (GDS) in stead of hippocampal volume. In this case the two conditional frequency distributions cannot be distinguished within the finite-sample uncertainty.

It must be remarked that it is exactly this kind of differences and uncertainties that clinical researchers often try to clumsily capture by using p-values. Several practitioners expressed relief and even awe at the possibility of visualizing the estimates of different sub-population statistics and even the uncertainties they carry because of finite sample size.

The quick analysis above was purely qualitative, but concrete numbers, such as quantiles and expected values, can be given instead. This becomes necessary when many variates are considered jointly and visualization is impossible. In such high-dimensional cases the package allows the user to compute any kind of distance between two frequency distributions (such as Hellinger or Kantorovich or Shannon-Jansen distance, or relative entropy) as well as its credibility intervals. The computation of the mutual information between any two sets of variates gives moreover a quantitative measure of their association that does not depend on assumptions such as linearity or gaussianity.

The ability to swap the "predictor" and "predictand" roles of any variates is illustrated in the plots

of fig. 2, which parallel those of fig. 1. The thicker lines show the probabilities

$$\mathrm{P}(\mathrm{cAD}{=}\mathrm{Y} \mid HV) \,, \qquad \mathrm{P}(\mathrm{cAD}{=}\mathrm{Y} \mid \mathrm{HV})$$

for various values of the HV variate, which becomes a predictor in this case. The clouds of thinner lines are samples of the corresponding probability distributions of the frequency distributions. On the plot on the left, for instance, we see that among individuals from this population having hippocampal volume around 5.0, between 30% and 40% will convert to Alzheimer's. Among those with volume around 2.0, we can only say that between roughly 45% and 80% will convert; in this case the finite-sample size (few samples with this HV value) leads to a much larger uncertainty of the frequency estimates.

The **inferno** package can work with any combination of continuous, discrete ordinal, discrete nominal, and binary variates. Continuous variates can also be defined in bounded intervals, and can also have boundary values capable of finite probability mass, as it may happen with censoring. The package cannot handle variates with complex topology, such as images, or periodic variates.

3 Nonparametric representation

In Bayesian nonparametric population inference the posterior probability distribution is over the set of all possible frequency distributions. The mathematical representation of this set is therefore crucial. The package uses the ingenious representation of a distribution as a mixture of product kernels introduced by Dunson & Bhattacharya 2011. For instance, for variates X and Y the frequency distribution is written as

$$F(X,Y) = \sum_i w_i \, K_X(X \mid \xi_i) \, K_Y(Y \mid \upsilon_i) \tag{4}$$

where w_i are positive and normalized weights, K_X is a distribution for variate X depending on parameters ξ_i, and similarly for K_Y. The product is easily generalized to any number of variates. In principle the sum should be countably infinite, but as discussed in Ishwaran & Zarepour 2002a; 2002b and in Dunson & Bhattacharya 2011, it is possible to truncate it to a finite number of values if an appropriate prior distribution is used for the weights w_i. Thus a frequency distribution F is effectively represented – non-uniquely – by a large but finite set of parameters (w_i, ξ_i, υ_i).

The above representation by products of individual kernels has been somewhat deprecated in more recent works (Wade et al., 2014b,a) which, however, consider inference problems where variates have clear predictor or predictand roles. As previously discussed, in many research fields such roles have no clear a-priori division and are quite dynamic. A representation that easily allow to swap the two role of variates, without overemphasizing either, is therefore most appropriate. The representation (4) leads to very simple and symmetric analytical expressions for the conditional of X given Y and vice versa, as well as any marginal:

$$\begin{aligned}
F(X \mid Y) &= \sum_i \frac{w_i}{\sum_j w_j K_Y(Y \mid \upsilon_j)} \, K_X(X \mid \xi_i) \, K_Y(Y \mid \upsilon_i) \\
F(Y \mid X) &= \sum_i \frac{w_i}{\sum_j w_j K_X(X \mid \xi_j)} \, K_Y(Y \mid \upsilon_i) \, K_X(X \mid \xi_i) \\
F(X) &= \sum_i w_i \, K_X(X \mid \xi_i) \,.
\end{aligned}$$

This posterior is represented by samples of such distributions, obtained by Markov-chain Monte Carlo.

[bayes base rate]

(Walker, 2010)

References

J.-M. Bernardo and A. F. Smith. *Bayesian Theory*. Wiley series in probability and mathematical statistics. Wiley, New York, repr. edition, 2000. doi:10.1002/9780470316870. First publ. 1994.

J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors. *Bayesian Statistics 9*. Oxford University Press, Oxford, 2011. doi:10.1093/acprof:oso/9780199694587.001.0001.

D. Draper. Assessment and propagation of model uncertainty. *J. R. Stat. Soc. B*, 57(1):45–70, 1995. doi:10.1111/j.2517-6161.1995.tb02015.x. See also discussion and reply in Spiegelhalter et al. (1995).

D. B. Dunson and A. Bhattacharya. Nonparametric Bayes regression and classification through mixtures of product kernels. In *Bernardo et al. (2011)*, pages 145–158. 2011. doi:10.1093/acprof:oso/9780199694587.003.0005.

N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors. *Bayesian Nonparametrics*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, 2010. doi:10.1017/CBO9780511802478.

M. G. M. Hunink, M. C. Weinstein, E. Wittenberg, M. F. Drummond, J. S. Pliskin, J. B. Wong, and P. P. Glasziou. *Decision Making in Health and Medicine: Integrating Evidence and Values*. Cambridge University Press, Cambridge, 2 edition, 2014. doi:10.1017/CBO9781139506779. First publ. 2001.

H. Ishwaran and M. Zarepour. Dirichlet prior sieves in finite normal mixtures. *Stat. Sinica*, 12(3): 941–963, 2002a. http://www3.stat.sinica.edu.tw/statistica/J12n3/j12n316/j12n316.htm.

H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the Dirichlet process. *Can. J. Stat.*, 30(2):269–283, 2002b. doi:10.2307/3315951.

D. V. Lindley and M. R. Novick. The role of exchangeability in inference. *Ann. Stat.*, 9(1):45–58, 1981. doi:10.1214/aos/1176345331.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, version 7.2 (4th pr.) edition, 2005. https://www.inference.org.uk/itila/book.html. First publ. 1995. See also video lectures at https://videolectures.net/events/course_information_theory_pattern_recognition.

P. Mana, P. G. Luca, I. Rye, A. Vik, M. Kociński, A. Lundervold, A. J. Lundervold, and A. S. Lundervold. Personalized prognosis & treatment using an optimal predictor machine: An example study on conversion from mild cognitive impairment to Alzheimer's disease, 2023. doi10.31219/osf.io/8nr56.

H. C. Sox, M. C. Higgins, D. K. Owens, and G. S. Schmidler. *Medical Decision Making*. Wiley, New York, 3 edition, 2024. doi:10.1002/9781119627876. First publ. 1988.

D. J. Spiegelhalter, A. P. Grieve, D. V. Lindley, D. V. Lindley, J. B. Copas, A. J. G. Cairns, C. Chatfield, G. Box, D. Cox, J. W. Tukey, A. E. Raftery, M. Aitkin, G. A. Barnard, P. Donnelly, A. S. C. Ehrenberg, A. E. Gelfand, B. K. Mallick, A. Gelman, X.-L. Meng, C. A. Glasbey, G. J. Gibson, R. Glendinning, U. Hjorth, R. E. Kass, L. Wasserman, M. Lavine, D. Madigan, B. M. Pötscher, J. W. Pratt, A. F. de Vos, B. J. Worton, A. Zellner, and D. Draper. Discussion of the paper by Draper. *J. R. Stat. Soc. B*, 57(1):71–97, 1995. doi:10.1111/j.2517-6161.1995.tb02016.x. See Draper (1995).

S. Wade, D. B. Dunson, S. Petrone, and L. Trippa. Improving prediction from Dirichlet process mixtures via enrichment. *J. Mach. Learn. Res.*, 15(30):1041–1071, 2014a. http://jmlr.org/papers/v15/wade14a.html.

S. Wade, S. G. Walker, and S. Petrone. A predictive study of Dirichlet process mixture models for curve fitting. *Scand. J. Stat.*, 41(3):580–605, 2014b. doi:10.1111/sjos.12047.

S. G. Walker. Bayesian nonparametric methods: motivation and ideas. In *Hjort et al. (2010)*, chapter 1, pages 22–34. 2010. doi:10.1017/CBO9780511802478.002.