

## PREDICTIVE INFERENCE AND DISCONTINUITIES

ELJA ARJAS

*Rolf Nevanlinna Institute, University of Helsinki, Helsinki, Finland*

*(Received February 1999; In final form March 2000)*

The purpose of this note is to study the consequences, mainly in the form of simple examples, which the fundamental ideas of deFinetti on predictive inference and exchangeable random variables have in the context of reliability problems. In particular, the role of unpredictable observations, or innovations, is related to discontinuities in the process of learning from the data. It is argued that probabilistic quantification of subjective uncertainties is an intrinsic element of realistic modeling. The presentation is mainly of an expository nature.

*Keywords:* Reliability; Exchangeability; Subjective probability; Bayes' formula; Causality; Minimal repair

### 1 INTRODUCTION

In important application areas of probability theory such as reliability, a widely accepted idea has been that the modeling can be done “just following the generally accepted axioms of probability”, without caring too much about how those probabilities are to be interpreted in the context. Perhaps it is thought that, in the relatively rare cases when the analyst's adherence to alternative interpretations of probability has a strong subjective element, the differences become only explicit if one actually carries on from probabilistic modeling to statistical data analysis and estimation. Statistical inference is then seen as something which can be thought of separately, and which is not directly influencing the way in which the probabilistic formulations are to be made.

Although there is some truth behind such an “agnostic” view, its followers may in practice be tricked into probability models and constructions which have little to do with the intuition that originally led to the problem formulation. (Some examples are considered in sections 3 and 4.) In particular, if one thinks of probability in subjectivist terms, then it becomes an intrinsically dynamic process of assessments, where the subject's uncertainties are continuously updated on the basis of observed data. Unpredictable, surprising observations are then reflected as discontinuities in the corresponding learning process. This can be contrasted with the “classical” view, that probabilities are mathematical expressions of certain properties of the considered objects, which then manifest themselves as regularities of relative frequencies in (often, only hypothetical) long series of repeated experiments or trials. This image of probability is a more static one, as the observer's learning process based on the

accumulating evidence contained in the observed data has been distanced away from the process of probabilistic modeling, into a set of statistical procedures that are to be executed separately – if at all.

To make the starting point almost trivial, consider first a single unit, say, a light bulb which is put on test for survival. Suppose “you” make initially your own subjective assessment of its future life length  $T$ , expressing it in the form of a probability distribution in the standard way:  $F(t) = \Pr(T \leq t)$ ,  $t > 0$ . For simplicity, take  $F$  to be absolutely continuous, with even the density  $f$  continuous. Suppose then that when monitoring the life of this light bulb, the only relevant additional information that will become available from the monitoring is whether it is still alive. At time  $u$  we can then decide by direct observation which of the two events  $\{T \leq u\}$  and  $\{T > u\}$  has happened, and in the former case we have actually come to know the value of  $T$ . Therefore, if  $u$  time units have passed without a failure, the original assessment will be updated into

$$F(t | u) = [F(t) - F(u)][1 - F(u)]^{-1} = \Pr(T \leq t | T > u), \quad 0 \leq u < t.$$

At the time of failure, monitoring obviously stops being informative.

Consider then the simple prediction problem in which one tries to estimate from the observed information the value of  $g(T)$ , where  $g$  is some test function of interest. We can write

$$M_u = E(g(T) | H_u) = \begin{cases} E(g(T) | T > u) = \int g(t) F(dt | u) & \text{on } \{T > u\} \\ g(T) & \text{on } \{T \leq u\} \end{cases} \quad (1.1)$$

where  $H_u$  is the history arising from the monitoring up to time  $u$ . The sample path of the process  $M = (M_u)$  starts from the value  $M_0 = E(g(T)) = \int g(t) F(dt)$ . If  $g$  is increasing, also  $M_u$  increases up to the point  $u = T$ , at which it falls down to the value  $E(g(T) | H_T) = g(T)$  and stays there from then on. This is illustrated in the case  $g(t) \equiv t$  in the following figure.

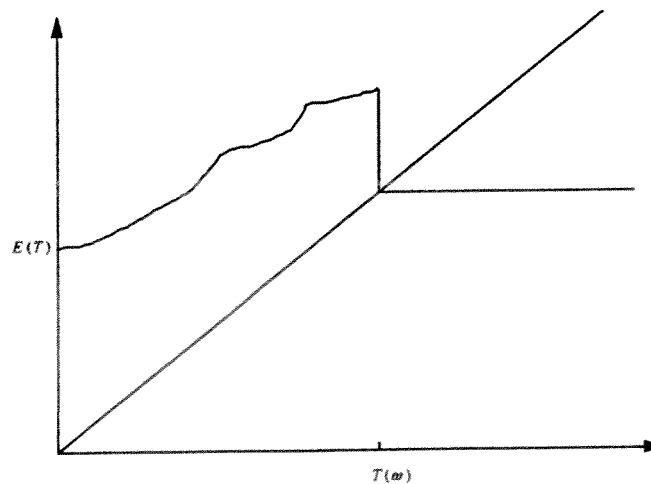


FIGURE 1 Typical sample path of the process  $(M_u)$  when considering a single life length  $T$ .

It is well known, and indeed easy to verify directly that, viewed as a stochastic process,  $(M_u)$  is a  $(Pr, H_u)$ -martingale. Its dynamics, representing updating of expected values of  $g$  as more information is becoming available, can be written in the form

$$M_u = M_0 + \int_0^u C_s [dN_s - \lambda_s ds], \quad u \geq 0, \quad (1.2)$$

where the process  $N_u = 1_{\{T \leq u\}}$ ,  $u \geq 0$ , counts “one” at time  $T$ , where  $\lambda_u = 1_{\{T \geq u\}} \cdot \lambda^F(u) = (1 - N_{u-}) \lambda^F(u)$  with  $\lambda^F(u) = f(u) [1 - F(u)]^{-1}$  is the corresponding  $(Pr, H_u)$ -intensity, and  $C_u = g(u) - E(g(T) | T \geq u)$  is called the innovation gain process [1]. Note that, if the test function  $g$  is chosen to be the identity,  $g(t) \equiv t$ , then  $-C_u = E(T - u | T \geq u)$  is the *expected residual lifetime* at times  $u$  before failure. Once  $F$  has been assigned, the only “randomness” in the sample path of  $(M_u)$  is in when  $T$  will happen, and since  $F$  was assumed continuous, the only discontinuity in the sample path of  $(M_u)$  is the jump downwards, of size  $|C_T|$ , at time  $u = T$ . This is the size of the disappointment experienced at the time of death: the expected residual life is suddenly becoming zero. Before time  $T$ , the path of  $(M_u)$  moves up “deterministically” (given  $F$ ) according to the derivative  $|C_u| \lambda^F(u)$ .

## 2 INFINITE EXCHANGEABILITY AND DISCONTINUITIES

This analysis becomes more interesting if we consider  $n$  ( $n > 1$ ) light bulbs instead of only one, with respective lifetimes  $T_1, T_2, \dots, T_n$ . Suppose that “you” have no reason for assigning probabilities to these  $n$  lifetimes differently. Therefore, it seems logical to assign originally the same distribution  $F$  to each of them. Consider the very simplest case in which the bulbs, while on test, are not in any physical contact with each other. Therefore, if one of them fails, those still on test will continue their life under conditions which are indistinguishable from the (counterfactual) situation in which there had been no failures.

In classical statistical modeling, such physical independence is thought to imply statistical independence. Under such an “i.i.d.” scheme, the joint distribution of the  $n$  lifetimes can be simply constructed as the product measure of the  $n$  marginals. Using that probability measure, we could formally repeat the above single light bulb analysis for each one of them without change, and without caring about what our monitoring reveals about the others. But is such independence really consistent with the natural goal that probability should be applicable as a person’s quantification of his uncertainty regarding future observables, that is, things that he has not (yet) observed or otherwise found out? For example, consider a situation in which “you” had originally put  $n = 100$  similar looking light bulbs on test, having then assigned the same distribution  $F$  to each of their lifetimes. If you then see several of them fail in a rapid succession soon after the test was started, should you still keep your original predictions concerning the surviving ones as if nothing had happened, only updating the probabilities according to the deterministic rule (1.1)?

No reasonable person, it seems, should be so stubborn as to remain uninfluenced by such new information. However, the ways in which the new information is going to be incorporated will in practice depend strongly on the adopted statistical paradigm. The solution offered by classical statistics, typically, is to choose some parametric model  $\{F_\theta, \theta \in \Theta\}$  for the lifetimes, where the parameter  $\theta$  is thought in reality to have some “fixed but unknown” value corresponding to an infinite population of light bulbs, and  $F_\theta$  is interpreted

in terms of relative frequencies in this population. A “plug-in” prediction based on this model and a point estimate of  $\theta$  is then the commonly offered solution, combined with a confidence band to settle the uncertainty, in frequentist terms, regarding the “true” value of  $\theta$ . When following this statistical procedure, one is adopting steps which are not necessary consequences of the axioms of probability theory. This differs in a fundamental way from the approach used in predictive inference, which builds largely on the notion of infinite exchangeability/extendibility (e.g., [2, 3]).

#### DEFINITION

- (i) *The joint distribution of random variables  $X_1, X_2, \dots, X_n$  is exchangeable if it remains the same regardless of how these random variables are permuted; in such a case also the variables  $X_1, X_2, \dots, X_n$  are called exchangeable.*
- (ii) *An exchangeable distribution  $F_n$  of  $n$  random variables is  $N$ -extendible if there is an exchangeable distribution  $F_N$  of  $N$  random variables, with  $N > n$ , such that  $F_n$  is a marginal of  $F_N$ , and it is infinitely extendible if it is  $N$ -extendible for any  $N > n$ .*

The idea behind this definition is that if one decides to consider (jointly) some specific number  $n$  individuals or objects one can always think of having selected them from a larger collection of  $N$  such objects, and the joint probability of those selected will not depend on either how much larger  $N$  is than  $n$ , nor on what particular subset of size  $n$  was selected.

Although this exchangeability/extendibility condition clearly says something about symmetry and independence, it is weaker than the standard i.i.d. assumption. In fact, the celebrated representation theorem deFinetti (e.g. [2, 3]) says that, under the infinite exchangeability/extendibility condition, one can always express the joint distribution of  $(T_1, T_2, \dots, T_N)$  in the mixture form

$$Pr(T_1 \leq t_1, \dots, T_N \leq t_N) = \int \prod_{1 \leq i \leq n} F(t_i) \pi(dF), \quad (2.1)$$

where  $\pi$  is a (probability) measure on the collection of distribution functions  $F$  on the positive half-line.

This shows that, under the exchangeability/extendibility assumption, the joint distribution of  $(T_1, T_2, \dots, T_n)$  can be expressed as a “prior predictive distribution”, where  $\pi$  takes the role of a prior,  $F$  is a parameter, and the likelihood corresponds to independent random sampling from  $F$ . Viewed formally as a simulation trial, one could produce a sample of size  $n$  according to  $Pr$  by considering a simple hierarchical (“doubly stochastic”) model: first realizing a distribution  $F$  according to the measure  $\pi$ , and then drawing an i.i.d. sample of  $n$  observations from this  $F$ . In the special case in which  $\pi$  has all its mass on a single distribution  $F^*$ , say, that is,  $\pi = \delta_{F^*}$ ,  $Pr$  will simply correspond to i.i.d. random sampling from  $F^*$ . But one should not be carried away by these frequentist interpretations! The starting point here was the probability  $Pr$  which was given a subjectivist interpretation; the representation (2.1) is only a mathematical theorem and it does not say anything about how  $\pi$  and  $F$  should be interpreted.

The real value of the (infinite) exchangeability/extendibility postulate is in its normative character: once it is accepted, then the way in which predictions based on observed data are to be updated becomes unique. All one needs to do is to compute the corresponding conditional probabilities according to (2.1).

As a first example of this, let us consider the notion of hazard rate or intensity. The standard way would be to start from the distribution  $F$  of a single light bulb, defining then

$$\lambda_u^F(i) = (1 - N_{u-}(i)) \lambda^F(u), \quad i = 1, \dots, n, \quad (2.2)$$

where  $N_u(i) = 1_{\{T_i \leq u\}}$  and  $\lambda^F(u) = f(u)[1 - F(u)]^{-1}$  is the hazard rate corresponding to  $F$ . However, this would mean ignoring the information contained in what has happened to the other light bulbs. Therefore, at time  $u$ , the joint distribution  $Pr$  of all  $n$  lifetimes should be conditioned on the observed pre  $u$  failure history  $H_{u-}$  (see, e.g., [3, 4]). All such histories can be expressed by simply listing the total number  $N_u = \sum_i N_u(i)$  of failed light bulbs up to time  $u$ , their indexes, and the precise times at which they failed. Because of the assumed exchangeability and the symmetry of the test function, however, the actual identities (and therefore indexes) of the light bulbs are irrelevant for the prediction.

Consider first the hazard assessment which you would make on any one of those light bulbs which have not already failed by time  $u$ , say the  $i$ th, based on the knowledge  $H_{u-}$  of failures strictly before time  $u$ . This corresponds to assessing the probability of the event that it will actually fail in a short interval from  $u$  to  $u + du$ . A simple calculation shows that this “infinitesimal” probability can be written as  $\lambda_u(i)du$ , where

$$\lambda_u(i) = \int \lambda_u^F(i) \pi_{u-}(dF), \quad (2.3)$$

and where

$$\begin{aligned} \pi_{u-}(dF) &\propto \pi(dF) \prod_{\{i: T_{(i)} < u\}} f(T_{(i)}) \times \prod_{\{j: T_{(j)} \geq u\}} [1 - F(u)] \\ &= \pi(dF) \prod_{i \leq N_{u-}} f(T_{(i)}) \times [1 - F(u)]^{n - N_{u-}} \end{aligned} \quad (2.4)$$

is the posterior distribution of  $F$  updated according to the observed failure data  $H_{u-}$ . Thus, apart from the observed “at risk” indicator  $1 - N_{u-}(i)$ ,  $\lambda_u(i)$  can be viewed as an expectation of the hazard rates  $\lambda^F(u)$  with respect to the posterior  $\pi_{u-}$  based on the data available just before time  $u$ . The role of the deFinetti representation here is in that it provides the simple rule according to which  $\pi_{u-}$  is to be computed from the “prior”  $\pi$  and the “data”  $H_{u-}$ : it is simply Bayes’ formula applied on the distributions  $F$ . It therefore seems appropriate to call  $\lambda_u(i)$  *predictive intensity*.

In particular, there is an instantaneous update in  $\pi_u$  at every failure time, and hence in the value of  $\lambda_u(i)$  in case the  $i$ th light bulb was one of those still at risk. This is of course trivial if the failing light bulb happens to be the  $i$ th, because its “own” failure intensity must clearly be zero after it has already failed. However, also the probability assessment of the future lives of those that survived will undergo a sudden change when one of their companion light bulbs fails. This is in spite of the fact that, as we have assumed, no physical damage was incurred on the survivors by these failures, and in this sense there is no causal dependence between the lifetimes of the light bulbs. Nevertheless, their predictive intensities  $\lambda_u(i)$  change abruptly at these points, resulting in discontinuities in the corresponding sample paths. Thus, starting from an absolutely continuous distribution  $F$  and the corresponding continuous hazard rate  $\lambda^F$ , we have arrived at predictive intensities which are left continuous but make jumps in an unpredictable manner.

The same reasoning carries over from hazard assessment of individual light bulbs to more general expectations of the form  $M_u = E(g(T_1, \dots, T_n) | H_u)$ , where  $g$  is a test function. Because of the assumed exchangeability between the light bulbs, we assume that  $g$  is symmetric in its arguments so that  $g(T_1, \dots, T_n) = g(T_{(1)}, \dots, T_{(n)})$ , where  $T_{(1)}, \dots, T_{(n)}$  are the order statistics of  $T_1, \dots, T_n$ . Intuitively, we can see that the information obtained from monitoring the light bulbs will become useful on two different levels. First, knowledge of  $H_u$  by time  $u$  implies that the first  $N_u = \sum_i N_u(i)$  coordinates of the vector  $(T_{(1)}, \dots, T_{(n)})$  are actually known, with values at most equal to  $u$ , while the remaining and still unobserved coordinates must have values which exceed  $u$ . Second, and more interestingly, this “first level” knowledge can be seen as data which carries information about the “parameter”  $F$ . Having once assumed the exchangeability condition, we can verify by a simple computation that the conditional independence structure of the representation (2.3) persists, although it now of course can be applied to only those light bulbs which so far did not fail. Now the mixing distribution  $\pi$  will be changed into  $\pi_u$ , the a posteriori distribution of  $F$  arising from “data”  $H_u$ . For those light bulbs which did fail by time  $u$ , the conditional distribution is naturally the point mass probability at the observed value  $(T_{(1)}, \dots, T_{(N_u)})$ , and for those that didn't, the conditional density, given  $F$  and survival up to time  $u$ , will be equal to  $f(t)[1 - F(u)]^{-1}$ ,  $t > u$ . Thus the above expectation can be written as

$$\begin{aligned} M_u &= E(g(T_{(1)}, \dots, T_{(n)}) | H_u) \\ &= \int g(T_{(1)}, \dots, T_{(N_u)}, t_{N_u+1}, \dots, t_n) \prod_{i=N_u+1}^n f(t_i)[1 - F(u)]^{-(n-N_u)} \pi_u(dF) dt_{N_u+1} \dots dt_n, \end{aligned} \tag{2.5}$$

where the integration is jointly over  $F$  and  $(t_{N_u+1}, \dots, t_n)$ .

As  $u$  increases, more failure times become known, and these known values are imputed directly as arguments into the function  $g$ . In this trivial sense, although the lifetimes  $T_1, \dots, T_n$  are not causally dependent on each other, they will all have a direct causal influence on the ultimate value of the test function  $g$ . After  $T_{(n)}$  the value of  $M_u$  remains of course fixed since ultimately “all lights are out” and  $g(T_{(1)}, \dots, T_{(n)})$  is known to the observer.

It is of some interest, however, to study somewhat more closely the jump behavior of the process  $M = (M_u)$ . Just as in the case of a single light bulb,  $M$  is a martingale with respect to the observed “internal” histories  $(H_u)$ , and therefore an integral representation analogous to (1.2) will hold. A straightforward computation shows that this representation is retained precisely in the same form as before if we only give the notations  $N_u = \sum_i N_u(i)$  and  $\lambda_u = \sum_i \lambda_u(i)$  their new meanings arising from considering  $n$  light bulbs instead of only one, and redefine the innovation gain process  $C_u$ . Intuitively speaking, once  $Pr$  has been specified, the only “randomness” in this example, and therefore also in the predictions  $(M_u)$ , is in the failure times  $T_{(1)}, \dots, T_{(n)}$ . Now the sample paths of  $(C_u)$  and  $(\lambda_u)$  are determined on each interval  $(T_{(i-1)}, T_{(i)})$  by the failure times  $T_{(1)}, \dots, T_{(i-1)}$  strictly before that interval begins, whereas  $(M_u)$  is similarly deterministic on each  $[T_{(i-1)}, T_{(i)})$ . The intuitive meaning of  $C_u$  is that it tells how  $M_{u-}$  is to be updated into  $M_u$  should there be a failure time at  $u$ .

It is actually immediately clear from (2.3), (2.4) and (2.5) what should happen to  $(M_u)$  at times  $u$  which coincide with one of the failure times  $T_{(i)}$ . At such time points  $N_u = N_{u-} + 1$ ,  $M_u = M_{u-} + C_u$ , and the history  $H_{u-}$  consisting of failure times  $T_{(1)}, \dots, T_{(N_{u-})}$  will be amended by a new failure time  $T_{(N_u)}$  into a history which we can conveniently denote by  $H_{u-} + \{u\}$ . Corresponding to this sudden change in the available data,

the left limit  $M_{u-} = E(g(T_{(1)}, \dots, T_{(n)}) | H_{u-})$  will be updated into the new value  $M_u = E(g(T_{(1)}, \dots, T_{(n)}) | H_{u-} + \{u\})$ . Therefore  $C_u$  can be viewed as the update in the value of  $M$  which is inflicted by a new failure time in case it occurs at time  $u$ . The expected value  $M_u$  can be computed from the expression (2.5), where

$$\pi_u(dF) \propto \pi(dF) \prod_{\{i: T_{(i)} \leq u\}} f(T_{(i)}) \times \prod_{\{i: T_{(i)} > u\}} [1 - F(u)] = \pi(dF) \prod_{i \leq N_u} f(T_{(i)}) \times [1 - F(u)]^{n - N_u}. \quad (2.6)$$

By the same token,

$$\begin{aligned} M_{u-} &= E(g(T_{(1)}, \dots, T_{(n)}) | H_{u-}) \\ &= \int g(T_{(1)}, \dots, T_{(N_{u-})}, t_{N_{u-}+1}, \dots, t_n) \prod_{i=N_{u-}+1}^n f(t_i) [1 - F(u)]^{-(n - N_{u-})} \\ &\quad \pi_{u-}(dF) dt_{N_{u-}+1} \dots dt_n, \end{aligned} \quad (2.7)$$

where  $\pi_{u-}$  was given in (2.4). When  $u$  is a failure time, we can quickly verify that

$$\pi_u(dF) \propto \pi_{u-}(dF) (n - N_{u-}) \lambda^F(u) \propto \pi_{u-}(dF) \lambda^F(u), \quad (2.8)$$

where the normalizing constants are given, respectively, by the observed overall predictive intensity  $\lambda_u$  and by the observed failure intensity  $\lambda_u^*$  of a generic light bulb \* still at risk just before time  $u$ . Thus Bayes' formula comes into play once again:  $\pi_{u-}$  is the prior,  $\pi_u$  is the posterior, and the likelihood arising from the new failure at  $u$  depends on whether we simply count the number of failures or also keep track on which particular light bulb it is that happens to fail. Because of the assumed symmetry, the difference is reflected only in the "constant" proportionality factor  $(n - N_{u-})$ , and therefore the results will be the same.

Let us now go back to the connection between the infinite exchangeability/extendibility condition and the familiar "i.i.d." assumption. As was emphasized above, the starting point in the former was a probability  $Pr$  satisfying the postulate of the Definition, which then gave the distribution  $F$  a role in which it was treated like a random element in a suitably defined function space. According to the deFinetti representation, data described by  $Pr$  can then be thought of as being "conditionally i.i.d., given the distribution  $F$ ". A natural question suggested by this is therefore: suppose "you" had a way of obtaining arbitrarily large samples of lifetimes  $T_i$  for which you would assign a probability  $Pr$  satisfying the postulate of the Definition. Would there then be some way of finding "the correct  $F$ "?

In fact, the question is somewhat ill-posed, because the basic postulate of exchangeability makes no reference to a data generating mechanism and a corresponding "correct  $F$ ". Instead, it has consequences regarding what you could expect to see in yet unobserved samples. An important implication of the exchangeability assumption, a Glivenko - Cantelli type theorem (see e.g. [3]) states that "you" can be certain (in the sense that you would under  $Pr$  assign probability one to this event) that, as the sample size  $n$  increases, the empirical distribution functions determined from such samples converge to a limiting cumulative distribution. This limit, which is random in the sense that it will always remain unknown to a real observer who only has a finite sample, is then taking the role of an unknown parameter in classical statistical inference. Specifying the value of that parameter, say  $F^*$ , corresponds

then to “your” pretension that you have already seen infinitely many of the exchangeable lifetimes, and have therefore come to know the limit of the empirical distribution functions.

Under such hypothetical circumstances, observing yet a few more lifetimes would, relatively speaking, contribute nothing new to your ability to predict others in the future. Therefore, hazard assessments concerning such future lifetimes could be made directly on the basis of  $F^*$  determined from the infinite sample. In particular, the innovation gains we have considered above would be identically equal to zero. This is the i.i.d. assumption, which in the present formulation corresponds to choosing  $\pi$ , and consequently all the  $\pi_u$ , to be point mass distributions at a particular  $F^*$ .

*Remark* Since the above reasoning does not make any reference to the concepts of an infinite population from which the data are drawn or a corresponding data generating mechanism, the classical consistency question, that is, whether following a particular statistical procedure will lead one to finding the “correct  $F^*$ ”, has no real content. However, making the extra postulate that the observed lifetimes are generated by “Nature” as an i.i.d. sequence of random variables from an unknown distribution  $F^*$ , this  $F^*$  can be recaptured from data asymptotically, by considering the empirical distribution functions. Thus, if Nature would indeed play such a game with you, and it would be possible to continue this game infinitely long, you would ultimately be able, from infinite data, to reveal Nature’s  $F^*$  with probability one by following the simple procedure of always computing the empirical distribution from the data that had accumulated so far. The difference between being a predictivist Bayesian or a frequentist is here in that, for a frequentist “probability one” refers to the data generating mechanism which he assumes will exist, whereas for a Bayesian it simply expresses the fact that he is certain of the event.

### 3 DISCONTINUITIES AND CAUSAL REASONING

With this we have covered, in the technically simple case of continuous densities, the situation in which the infinite exchangeability/extendibility assumption and the consequent deFinetti representation (2.1) hold. Our next question is therefore: what can be said if this assumption is not made? What properties and statements will still remain true? A related issue, which is important from the point of view of probabilistic modeling, is what kind of system structures in a reliability context would make that postulate unnatural?

Let us consider the last question first. In the light bulb example, we could quite easily envision a situation in which more and more light bulbs would be tested in parallel, without that this would in any way affect those  $n$  which are really being considered. Moreover, any collection of  $k$  light bulbs could be expected to behave in a similar fashion regardless of what their indexes were. But the situation is no longer the same if we have an actual device or “system” consisting of several parts. There is of course no hope of any exchangeability property if the parts are of different designs, performing different functions, etc. But even if they themselves are considered to be similar enough so that one would naturally assign the same joint distribution to their lifetimes regardless of the order in which they are indexed, and therefore satisfy the exchangeability, it is hard to see that any real devices performing an actual function would satisfy the infinite extendibility condition. For example, considering originally a 4-cylinder car engine, could anybody realistically imagine adding an arbitrarily large number of new cylinders to such an engine, and then say that, for each  $k = 1, 2, \dots$ , all “ $k$ -cylinder marginal engines” could be expected to function in a similar fashion regardless



of what particular cylinders were chosen, and therefore be assigned the same probabilistic description? It is actually hard to imagine what such marginal engines would be like, and what bearing their contemplated performance or reliability would have from the point of view of the 4-cylinder engine actually considered.

The same difficulty persists even in idealistic mathematical reliability models, such as parallel or series systems consisting of similar parts or components, where the symmetry between these parts is by direct assumption. Under such circumstances it is not difficult to imagine that more and more similar parts could be added forming then, together with the original ones, a larger series or parallel system. What is not realistic, however, is that we could consider any  $k$  parts “marginally”, in isolation, expecting them to behave in a similar manner regardless of how many other similar parts there might be connected into the same system. For example, a series system stops working as soon as the first of its parts fails, after which it no longer makes sense to talk about the reliabilities or hazards of the other parts. It will therefore be crucial how many parts there are in the system altogether, and if that number changes, also all relevant probability assessments will change.

For an illustration, let us consider a parallel system consisting of  $n$  similar parts. This structure implies that (i) the system is in a working condition as long as at least one of its parts is working, and (ii) the hazards of all such parts are to be assessed symmetrically. Simplifying the assumptions still further, we rule out the possibility that more than one part would fail at precisely the same time. While still a simplistic description of any real technical system or device, this structure has an important new ingredient which was not present in the light bulb example: when a part of the parallel system fails, this will generally have a causal effect on the lifetimes of the parts that remain in an operational state. In concrete situations, such causal dependence can be thought to arise from one or more physical mechanisms present in the system and influencing the environment in which the surviving parts serve their function. For example, it would be reasonable to assume that a failure of a part, if it is not critical to the system and killing it right away, always increases the load on the remaining parts.

As argued above, we can no longer reasonably assume that the infinite exchangeability/extendibility condition stated in the Definition would hold for such a parallel system, and therefore cannot base our predictions on the powerful deFinetti representation. However, the martingale representation (1.2) is still valid, with its structure and logic virtually unchanged. Thus we can continue to make probabilistic assessments about the future of the system and of its parts, basing such assessments on the monitored failure data. Again, these part failures are all we can actually observe, and in this sense “all randomness” that can be settled directly from the data is in when they will happen. In particular, the predictive intensities ( $\lambda_u$ ) and the innovation gains ( $C_u$ ) will behave “deterministically” on random intervals of the form  $(T_{(i-1)}, T_{(i)}]$ , then only depending on the past failure times  $T_{(1)}, \dots, T_{(i-1)}$  strictly before that interval begins, whereas ( $M_u$ ) is similarly deterministic on each  $[T_{(i-1)}, T_{(i)})$ . The connection between the innovation gains ( $C_u$ ) and the martingale ( $M_u$ ) is also retained, in the sense that  $C_u$  is equal to the jump  $\Delta M_u = M_u - M_{u-}$  in the sample path of ( $M_u$ ) in case  $u$  happens to be the failure time of one of the parts at risk.

What will not remain the same, however, is the interpretation of such a discontinuity of ( $M_u$ ). In the light bulb example, it was an intrinsic property of the considered “system” that its part lifetimes were not physically influencing each other. All instantaneous updating of the predictive intensities was therefore thought to result from updating a “prior”  $\pi_{u-}$  into a “posterior”  $\pi_u$  if indeed  $u$  happened to be a failure time. Now that the deFinetti representation is no longer available, there is no norm about how the discontinuities of the prediction martingale should be determined. Moreover, the discontinuities can be seen as containing two elements: (i) a *causal effect* which, typically via some physical mechanisms triggered

by a part failure start influencing the remaining lives of the parts which are still at risk, and just as in the light bulb example, (ii) an *information effect* coming from the knowledge that a failure occurred. (As a convention, we include as part of the information effect the direct functional effect on the considered test function  $g$  coming from the fact that one of the arguments becomes known when a part fails.) The total size of the discontinuity is then the sum of the causal effect and the information effect. An analogous distinction between “physical” and “psychological” effects was made by Lynn and Sinpurwalla [5] in connection with burn-in, and by Apostolakis and Wu [6] who talked about “stochastic” and “state-of-knowledge” dependence.

Although there is no norm for how the size of such discontinuities should be determined without assuming further model structure, it is nonetheless of some interest to ask whether there is some principle according to which discontinuities in the prediction martingale ( $M_u$ ) could be decomposed into their two components representing different types of effects. The following idea seems to apply here: Suppose that a part, instead of failing “naturally” by itself, is killed by an intervention from outside. Suppose also that this is done in a way which, from the point of view of the parts remaining at risk, is as if the death had been “natural”. As a consequence, the causal effect of a failure on the system can be assessed as being the same regardless of whether it was caused by an intervention or by a natural death. On the other hand, a failure induced by an intervention from outside should not contain an information effect. Thus, if we had a way of determining the causal effect of an outside intervention, this could be directly identified with the causal component of the discontinuity resulting from a natural death. This idea of separating two kinds of conditioning, depending on whether the sequence of events which led to the failure was viewed as being “endogenous” or “exogenous” to the considered process, appears to be analogous to the distinction between ordinary and “check” conditioning made by Pearl [7].

Unfortunately, making the conceptual distinction between causal and information effects does not help directly in quantifying or estimating their sizes in nontrivial concrete situations, with finite amounts of data. The only systematic way for approaching such a problem seems to be to set up a hierarchical Bayesian model describing the considered system probabilistically in an honest way, including the specification of the priors, and then do a posterior analysis based on observational and experimental data. In essence, this suggestion follows the program presented by Chen and Singpurwalla [8].

Hierarchical Bayesian models, unless they are firmly based on the idea of exchangeability, have a strong element of subjectivity coming from the use of expert judgment. This can be a virtue, but it can also lead to ad hoc formulations, and the latter aspect is likely to become more worrying as the degree of the difficulty of the substantive problem and the size of the model increase. In particular, in complex problems there is rarely an operational definition which would characterize the unobservables of the assumed model in a precise manner. Therefore the relevance of the conclusions made on the basis of such hierarchical Bayesian models will always depend on how convincing the assumed model structure is to “you”, or to whoever is asked to accept the underlying argumentation and the conclusion made.

#### 4 DISCUSSION

Reliability theory, in the way it is commonly practiced today, can be viewed as being a specialized study of lifetime distributions. In view of this, and particularly thinking about the practical goals of that theory in assessing risks which arise in the context of large technological systems, it would be desirable if these distributions could be given some reasonably

direct physical meaning. While the technical terminology which has been adopted actually reflects such an ambition, one may ask how well it is really justified. For example, *aging*, which in common language refers to certain irreversible complex physical (or physiological) processes taking place in the considered objects or individuals, is in reliability theory defined through some particular classes of lifetime distributions, characterized by inequalities which they are supposed to satisfy. This leads to a troubling confusion between actual aging and a selection mechanism acting in a heterogeneous population.

To provide another example where conventional reliability modeling and terminology do not appear to correspond to ordinary semantics of the English language, consider the notion of *minimal repair*. In words, minimal repair is defined as an operation by which a device (or its part) is after a failure reinstated back into the functional state in which it was just before it failed. Alternatively, minimal repair is said to be an operation by which the device is immediately after a failure replaced by another, which is similar in other respects except that it did not experience the failure in question. Although these definitions are sometimes distinguished from each other by calling the former “physical” and the latter “statistical” minimal repair [9, 10], they are usually treated as if they were equivalent. But are they the really?

For a person who has postulated exchangeability for a class of individual lifetimes, an observed failure carries two types of messages. First, as discussed in section 2, there is a collective effect on the entire exchangeable class: it creates an instantaneous update, a discontinuity, into the predictions concerning all similar devices still at risk. The representation theorem of deFinetti and Bayes’ formula together provide here a normative rule about how that update should be determined. Therefore, unless the empirical evidence from the exchangeable class is so extensive that one can actually claim to already know  $F^*$  with high precision, one should after a statistical minimal repair not make probabilistic predictions concerning the future of the members of that class, and in particular of the failed device itself, as if nothing at all had happened. Nevertheless, this is precisely how such minimal repair is usually modeled: the sequence of successive (minimally repaired) failures is treated as a Poisson process, with deterministic intensity equal to the hazard rate  $\lambda^F(u) = f(u)[1 - F(u)]^{-1}$  corresponding to an unknown lifetime distribution  $F$ , typically thought to represent relative failure frequencies in a population of similar devices. The discontinuities which are present in the intensity (2.3) have thereby disappeared.

Second, and perhaps more importantly, an observed failure as a rule carries a strong message about the particular device which failed: it was very likely already “weak” shortly before it did so, either because it had been weak from the time when it was new or because some in the context relevant aging process had advanced far enough to make it weak. Since physical minimal repair is only supposed to “resuscitate” a device back to life but not remove its weakness, this notion appears to be really quite different from statistical minimal repair. In all likelihood, a physically minimally repaired device will fail soon again so that its relevance as a repair operation may well be questioned. Thus the failure intensity should increase dramatically after a physical minimal repair, creating a discontinuity much larger than one would typically find in (2.3). Yet, physical minimal repair is traditionally modeled exactly as statistical minimal repair, in terms of a Poisson process with a deterministic intensity referring to a population of similar but still functioning devices. Heterogeneity in the quality of the devices and the resulting selection when they are in operative use are then ignored completely. This model may in some sense be “objective”, but it appears to have very little to do with realistic description of actual resuscitation or physical minimal repair. All realistic attempts to model such repair are likely to be much more difficult. No wonder, even sharply different opinions on the roles of objectivity and subjectivity have been expressed, see eg. [11, 12].

Physical minimal repair is an example of a common situation in which frequentist probabilities representing a collective are assigned as such to an individual member of that

collective, although there happens to be relevant additional information (here failure and a subsequent repair) that make this member to be a “special” one (cf. [13]). In general, I think that it is a serious mistake to interpret frequentist probabilities estimated from a collective as if they were physical characteristics of its individual members. Using a Poisson process model for physical minimal repair does exactly that. Clinical decision making in the context of treating individual patients is another area where “objective” nonadaptive application of population based frequentist probabilities as a norm can be directly misleading and harmful (see e.g. [14]). Predictive probabilities should be allowed to jump when important new information becomes available!

### References

- [1] Arjas, E. (1993). Information and reliability: a Bayesian perspective. In: Barlow, R. E., Clarotti, C. A. and Spizzichino, F. (Eds.), *Reliability and Decision Making*. Chapman & Hall, London, pp. 115–136.
- [2] deFinetti, B. (1975). *Theory of Probability*. Wiley, New York.
- [3] Spizzichino, F. (2001). *Subjective Probability Models for Lifetimes*. Chapman & Hall/CRC, Boca Raton.
- [4] Arjas, E. (1989). Survival models and martingale dynamics (with discussion). *Scand. J. Statist.*, **16**, 177–225.
- [5] Lynn, N. J. and Singpurwalla, N. D. (1997). Comment: “Burn-in” makes us feel good. *Statistical Science*, **12**, 13–19.
- [6] Apostolakis, G. E. and Wu, J. S. (1993). The interpretation of probability, deFinetti representation theorem, and their implications to the use of expert opinions in safety management. In: Barlow, R. E., Clarotti, C. A. and Spizzichino, F. (Eds.), *Reliability and Decision Making*. Chapman & Hall, London, pp. 311–322.
- [7] Pearl, J. (1995). Causal diagrams for experimental research (with discussion). *Biometrika*, **82**, 669–710.
- [8] Chen, J. and Singpurwalla, N. D. (1996). The notion of “composite reliability” and its hierarchical Bayes estimation. *J. Amer. Statist. Assoc.*, **91**, 1474–1484.
- [9] Bergman, B. (1985). On reliability theory and its applications (with discussion). *Scand. J. Statist.*, **12**, 1–41.
- [10] Arjas, E. and Norros, I. (1990). Should minimal repair depend on information? In: Block, H. W., Sampson, A. R. and Savits, T. H., (Eds.), *Topics in Statistical Dependence*, IMS Lecture Notes – Monograph Series, Vol. 16, pp. 5–12.
- [11] Evans, R. A. (1989). Bayes is for the birds. (Editorial). *IEEE Trans. Reliability*, **38**, 401.
- [12] Kaplan, S. (1990). Bayes is for eagles. (Editorial). *IEEE Trans. Reliability*, **39**, 130–131.
- [13] Dempster, A. P. (1998). Logicist statistics I. Models and modeling. *Statistical Science*, **13**, 248–276.
- [14] Klemola, U-M. and Norros, L. (1997). Analysis of the clinical behaviour of anaesthetists: Recognition of uncertainty as a basis for practice. *Med. Educ.*, **31**, 449–456.