Statistica Neerlandica (1997) Vol. 51, nr. 3, pp. 287–317

Conditioning as disintegration

J. T. Chang and D. Pollard*

Statistics Department, Yale University, Box 208290 Yale Station, New Haven, CT 06520, USA

Conditional probability distributions seem to have a bad reputation when it comes to rigorous treatment of conditioning. Technical arguments are published as manipulations of Radon–Nikodym derivatives, although we all secretly perform heuristic calculations using elementary definitions of conditional probabilities. In print, measurability and averaging properties substitute for intuitive ideas about random variables behaving like constants given particular conditioning information.

One way to engage in rigorous, guilt-free manipulation of conditional distributions is to treat them as disintegrating measures—families of probability measures concentrating on the level sets of a conditioning statistic. In this paper we present a little theory and a range of examples—from EM algorithms and the Neyman factorization, through Bayes theory and marginalization paradoxes—to suggest that disintegrations have both intuitive appeal and the rigor needed for many problems in mathematical statistics.

Key Words & Phrases: Conditional probability distributions, disintegrations, EM algorithm, sufficiency, Bayes theory, admissibility, marginalization paradoxes, Basu's theorem, exchangeability.

1 Introduction

In elementary probability courses one learns to calculate conditional probabilities by taking ratios, sometimes on little intervals that shrink to a point at the end of a proof. Conditional probability distributions are used and enjoyed freely, in restricted settings. In more advanced courses, where conditioning is placed on a rigorous measure-theoretic basis, one learns that real probabilists use Radon–Nikodym derivatives. One is warned that only in special cases can the conditional expectation $H_X(t) = \mathbb{P}(X \mid T = t)$ be treated rigorously as the expectation of the random variable X with respect to a probability measure $\mathbb{P}(\cdot \mid T = t)$ that concentrates on the set $\{T = t\}$. Instead, in the abstract Kolmogorov approach, $H_X(\cdot)$ is characterized up to almost-sure equivalence as the measurable function for which

$$\mathbb{P}[\{T \in B\}H_X(T)] = \mathbb{P}[\{T \in B\}X] \tag{1}$$

chang@stat.yale.edu, pollard@stat.yale.edu, world-wide web URL http://www.stat.yale.edu
* Supported by NSF Grants DMS-9102286 and DMS-9404180

for all measurable B. (Note that we are using linear functional notation for expectations, as explained at the end of this section.) The abstract approach has the virtue of making $\mathbb{P}(X \mid T = t)$ well defined (up to an almost sure equivalence) as a function of t whenever X is integrable. It has the disadvantage of sacrificing intuition to rigor.

Conditional probability distributions are clearly missed in some advanced work. Probabilists and statisticians often really do think in terms of conditional distributions, returning to them for private side calculations performed to get initial understanding of a problem. One first guesses the form of $H_X(t)$, perhaps with the help of an unjustified manipulation of the nonexistent probability measure $\mathbb{P}(\cdot \mid T = t)$, or by a hand-waving reduction to the discrete case. Then the proof reduces to a mechanical checking of the necessary measurability and averaging properties. Moreover, attempts to construct rigorous arguments using only elementary methods of conditioning can lead to the imposition of extraneous regularity conditions. Such attempts also lead to contortions, such as the introduction of unnecessary random variables and maps that transform the problem to a setting in which conditional densities may be calculated as ratios of joint to marginal densities on Euclidean spaces.

In this paper we discuss an approach to conditioning that combines the advantages of both the elementary and the abstract Kolmogorov approaches. We advocate the use of disintegrations, which are regular conditional distributions $\mathbb{P}(\cdot \mid T = t)$ that also satisfy natural concentration requirements of the form $\mathbb{P}\{T \neq t \mid T = t\} = 0$. We borrow the term "disintegration" from the French to emphasize the extra concentration property. The level of generality achievable by the disintegration approach to conditioning is much higher than with elementary methods. The extra requirements do sacrifice slight generality compared with the abstract Kolmogorov approach, but in the problems that we consider the generality is not missed. As compensation, arguments using disintegrations tend to look and feel much closer to the elementary arguments; by aiming for slightly less generality, we get to make stronger statements that come closer to the way that we tend to think intuitively about conditioning.

Consider a typical example.

EXAMPLE 1. The intuitive definition of sufficiency says that a statistic T is sufficient for a family of probability measures $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ if the conditional distributions given T do not depend on θ. The elementary approach is based on conditional distributions, which work fine in the simplest discrete and absolutely continuous settings, but are typically abandoned in rigorous treatments that aim for any more generality. As LEHMANN (1959, page 18) noted, there are some "difficulties concerning the behavior of conditional probabilities" that make a precise analysis delicate.

As an example, suppose $\mathbb{P}_\theta$ is the uniform distribution on the square $[0, \theta]^2$, for an unknown positive θ. The coordinate maps X and Y are independent Uniform $[0, \theta]$ under $\mathbb{P}_\theta$. The maximum, M, of X and Y is a sufficient statistic. Given $M = m$, the

conditional distribution $\mathbb{P}_\theta(\cdot \mid M = m)$ is uniformly distributed around two edges where one of X or Y equals m and the other is smaller.

One could argue informally, by conditioning on $\{m \le M \le m + \delta\}$ and then letting δ tend to zero, to get the form of the conditional distribution. It is also easy to check the Radon–Nikodym property by direct calculation of probabilities, but we feel that it is helpful to be able to think of the conditional distribution concentrated around the two edges where $M = m$.

Frequently one sees sufficiency for this particular example demonstrated by an appeal to a factorization theorem for the joint density of X and Y. A diligent student might be dismayed to learn that the form of that theorem needed in the present simple case is beyond the scope of most texts—typically one is offered the proof for the simple, discrete version of the theorem, with a suggestion to read about the general case in the thorough text of LEHMANN (1959). Even there, one might suspect that the simple proof (page 19) for smooth continuous distributions might have small problems with the non-differentiability of the maximum function. To be really rigorous one seems forced to skip forward several sections (to page 46) to find Lehmann's treatment of the HALMOS and SAVAGE (1949) approach, based on Radon–Nikodym derivatives.

Is it really that complicated? See Example 6. □

It has long bothered us (and other authors, such as TJUR, 1974 and WINTER, 1979) that there should be such a wide gap between intuition and rigor in conditioning arguments. We feel that, in many statistical problems, manipulation of the conditional probability distribution is the most intuitive way to proceed. However, we mathematical statisticians are trained to treat such conditional distributions with great caution, being aware of the menagerie of nasty counterexamples—such as the Borel paradox—that warn one away from conditional distributions. Apparently such examples have left conditional distributions with a bad name. As KOLMOGOROV (1930, page 51) put it, "the concept of a conditional probability with regard to an isolated hypothesis whose probability equals 0 is inadmissible." There is a technical difficulty, but it does not require us to abandon the notion of a conditional distribution. We feel our profession may have overreacted to the difficulties of placing conditioning on a sound basis and, in so doing, given up too much of the power of intuition.

By way of a small amount of theory and a collection of illustrative examples, in this paper we present a case that disintegrations are easy to manipulate and that they recapture some of the intuition lost by the more abstract approach, allowing guilt-free manipulation of conditional distributions. Most of our mathematics is well known and well used in certain areas of probability theory, such as Markov process theory. The disintegration property is essentially the assertion of the Decomposition Theorem in Section 29.2 of LOÈVE (1978), or of Theorem 6 in Section 2.5 of LEHMANN (1959). (For further references see Section 5.) Nevertheless, it seems to us that the ideas are not as widely known or used as they should be, which is our reason for collecting together some of the facts we might easily have learned in graduate school,

but didn't. We suggest that the concept of disintegration should be part of the education of every young probabilist and mathematical statistician.

In Section 2 we outline some theory for disintegrations, which we apply to a collection of conditioning examples in Section 3. We would suggest that the reader might contemplate how one usually attacks these problems, before looking at our explanations. We were all too often surprised and embarrassed by how much difficulty we were having using traditional methods as a first pass on the problems during the drafting of the paper.

With some slight trepidation—we fear some readers might take fright at the absence of integral signs—we have chosen to use notation that we have found most convenient and most helpful to our understanding. We adopt linear functional notation for integrals, writing λf instead of $\int f d\lambda$ or $\int f(x)\lambda(dx)$. We also identify sets with their indicator functions: $\lambda(fA)$ instead of $\int f(x)1\{x \in A\}\lambda(dx)$. When we want to identify explicitly the dummy variable of integration—for example, when integrating a function of more than one variable—we do so by attaching a superscript to the measure: $\lambda^y f(x,y)$ is the same as $\int f(x,y)\lambda(dy)$.

We also adopt a slightly unusual notation for image measures. If T is a measurable function from $(\mathcal{X},\mathcal{A})$ into $(\mathcal{T},\mathcal{B})$, and if λ is a measure on $(\mathcal{X},\mathcal{A})$, we denote the *image measure* of λ under the map T by $T(\lambda)$, or simply $T\lambda$. It is defined by

$$(T\lambda)(g) = \lambda(g \circ T)$$

for nonnegative measurable functions g on $(\mathcal{T},\mathcal{B})$. If λ is a point mass at x_0 then $T\lambda$ is a point mass at Tx_0. If g is the indicator function of a set B then $g \circ T$ is the indicator function of the inverse image $T^{-1}B$, and $(T\lambda)B = \lambda(T^{-1}B)$. That is, our $T\lambda$ is the same as the measure sometimes denoted by λT^{-1}. If λ is a probability measure, $T\lambda$ is also called the *distribution* of T under λ.

2 What is a disintegration?

In the elementary approach to conditioning, there are two ways to calculate conditional distributions. In the discrete case everything reduces to ratios of probabilities. For continuous distributions on Euclidean spaces (that is, distributions absolutely continuous with respect to Lebesgue measure), with conditioning on the projection onto a coordinate space, one calculates conditional densities by dividing marginal densities into joint densities. Conditioning on other random variables (or vectors) presents some difficulty when contemplated in any generality. Special transformations under extra smoothness assumptions are needed to reduce the calculations to the special case.

In this section we will describe a method that covers both discrete and continuous cases with equal ease. The same formulae appear in all cases. But first let us consider the elementary discrete case in more detail.

For discrete random variables conditioning is straightforward, as long as we heed the admonition not to try to condition on events of probability zero. Suppose $\mathbb{P}$ is a

probability measure on $(\mathcal{X}, \mathcal{A})$. Suppose T takes values only in a finite subset $\mathcal{R}$ of $\mathcal{T}$, with $\mathbb{P}\{T = t\} > 0$ for all t in $\mathcal{R}$. The elementary definition has

$$\mathbb{P}(A \mid T = t) = \frac{\mathbb{P}A\{T = t\}}{\mathbb{P}\{T = t\}} \quad \text{for } A = \mathcal{A} \text{ and } t = \mathcal{R}$$

Following KOLMOGOROV (1933), we will also use the more compact $\mathbb{P}_t(A)$ for $\mathbb{P}(A \mid T = t)$. The elementary definition enjoys the following pleasant properties.

(a) $\mathbb{P}_t(\cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{A})$ for all $t \in \mathcal{R}$.

(b) The measure $\mathbb{P}_t$ concentrates on the set $\{T = t\}$:

$$\mathbb{P}_t\{T \neq t\} = \frac{\mathbb{P}\{T \neq t\}\{T = t\}}{\mathbb{P}\{T = t\}} = 0$$

(c) For $A \in \mathcal{A}$,

$$\mathbb{P}(A) = \sum_{t \in \mathcal{R}} \mathbb{P}\{T = t\}\mathbb{P}_t(A)$$

The decomposition in property (c) writes $\mathbb{P}$ as a weighted sum of the conditional probability measures $\mathbb{P}_t$ for t in $\mathcal{R}$, where the measure $\mathbb{P}_t$ concentrates on the level set $\{T = t\}$. Notice that $\mathbb{P}\{T = t\}$ is the mass placed by the image measure $T\mathbb{P}$ at the point t. In our notation, the averaging property in (c) is written

$$\mathbb{P}A = (T\mathbb{P})^t \mathbb{P}_t A$$

We would like with equal assurance to be able to talk about and work with conditional probabilities of the form $\mathbb{P}(A \mid T = t)$ for more general spaces $(\mathcal{X}, \mathcal{A})$ and maps T. The standard Kolmogorov definition of conditional expectation has an accounting problem: for each $A \in \mathcal{A}$ the measurable function $\mathbb{P}(A \mid T = t)$ is free to be defined arbitrarily on any set of probability zero, and as there are in general many events $A \in \mathcal{A}$, those sets of probability zero could accumulate into a nonnegligible set. Worse yet, however many events there may be, there are still more sequences of events. For each such disjoint sequence $A_1, A_2, \ldots$, we have relations like

$$\mathbb{P}(\cup A_n \mid T = t) = \Sigma\, \mathbb{P}(A_n \mid T = t) \tag{2}$$

holding, at least almost surely, in the sense that there exists a set $N \subseteq \mathcal{T}$ for which $\mathbb{P}\{T \in N\} = 0$ and for which (2) holds if $t \notin N$. The set N depends on the particular sequence $\{A_i\}$. Thus, the unpleasant prospect arises that there might be no t for which (2) holds simultaneously for all sequences $A_1, A_2, \ldots$ of disjoint events.

These considerations are the familiar motivation for introducing the concept of a regular conditional distribution. Under stronger assumptions than required for the existence of Kolmogorov's conditional expectations, one can choose appropriate versions of each $\mathbb{P}(A \mid T = t)$, as a function of t, to make $\mathbb{P}(\cdot \mid T = t)$ a probability measure for (almost) all t. The slightly stronger notation of a disintegration also requires $\mathbb{P}(\cdot \mid T = t)$ to concentrate on the set $\{T = t\}$.

For the general definition of a disintegration we will consider not just probability measures, but also measures (such as Lebesgue measure) that have infinite total mass. Let T be a measurable map from $(\mathcal{X}, \mathcal{A})$ into $(\mathcal{T}, \mathcal{B})$. Let λ be a sigma-finite measure on $\mathcal{A}$ and μ be a sigma-finite measure on $\mathcal{B}$. Here λ is the measure to be disintegrated and μ is often the image measure $T\lambda$, although, as we will see below, it is useful to admit other possibilities for μ, especially to cover cases where $T\lambda$ is not sigma-finite.

DEFINITION 1. *We say that λ has a disintegration $\{\lambda_t\}$ with respect to T and μ, or a (T, μ)-disintegration, if:*

(i) *λ_t is a sigma-finite measure on $\mathcal{A}$ concentrated on $\{T = t\}$, that is, $\lambda_t\{T \neq t\} = 0$, for μ-almost all t;*

and, for each nonnegative measurable f on $\mathcal{X}$:

(ii) *$t \mapsto \lambda_t f$ is measurable;*
(iii) *$\lambda f = \mu^t(\lambda_t f)$.*

We will refer to the $\{\lambda_t\}$ as the disintegrating measures and to μ as the mixing measure. We will also write $\lambda(\cdot \mid T = t)$ for $\lambda_t(\cdot)$ on occasion.

Requirement (i) is analogous to property (b) in the discrete case; requirement (iii) is the analog of (c) generalized to functions. As defined by DELLACHERIE and MEYER (1978, page 78) the disintegrating measures $\{\lambda_t\}$ are required to be probability measures, analogously to (a). However we find that it is better to hold that property in reserve, and allow more general disintegrating measures. (Purist disintegrators might prefer us to invent yet another name.) As we will soon demonstrate, the λ_t can be taken as probability measures if and only if the image measure $T\lambda$ is sigma-finite and we take μ to be that image measure. In that case we will speak of a T-disintegration, omitting explicit mention of μ.

When λ and (almost) all the λ_t are probability measures we will also refer to the disintegrating measures as *(regular) conditional distributions* or *(regular) conditional probabilities*; we will usually write $\mathbb{P}$ and $\mathbb{P}_t$, instead of λ and λ_t, in this case. If X is a $\mathbb{P}$-integrable random variable, its expecation with respect to $\mathbb{P}_t$ is then a version of the conditional expectation $\mathbb{P}(X \mid T = t)$. As shown in Section 6, the concentration property (i) for conditional probabilities is a simple consequence of (1) when the sigma-finite $\mathcal{B}$ contains all singleton sets and is countably generated. Thus a disintegration of a probability measures may be thought of as resulting from a careful selection of versions of the conditional expectations (in Kolmogorov's sense), in a way that eliminates awkward complications caused by uncountable families of negligible sets. Not surprisingly, as with many stochastic process problems involving uncountable families of random variables, we need some extra (topological) assumptions about the underlying spaces and maps to ensure existence of the disintegration.

It might appear that, as a proper probabilist or mathematical statistician, one should be interested only in the case where the disintegrating measures are

probabilities. However, then one could not recover as a special case of a general disintegration result the elementary formula for calculating a conditional density (with respect to Lebesgue measure) as a ratio of a joint density to a marginal density. It would also hamper the improper urges of Bayesians with their priors (see Example 9).

EXAMPLE 2. Suppose λ is a product of two sigma-finite measures, $\lambda = \nu \otimes \mu$, on a product space $\mathcal{S} \otimes \mathcal{T}$. Let T be the map that projects onto the $\mathcal{T}$ coordinate space. For example, λ might be Lebesgue measure on $\mathbb{R}^2$ and μ might be Lebesgue measure on the x-axis.

Think of $\mathcal{S} \otimes \{t\}$ as a copy of $\mathcal{S}$ imbedded into the product space, and let λ_t be ν living on that copy. With a mild abuse of notation we will write $\lambda_t = \nu$. (More formally, let λ_t be the image of λ under the map $s \mapsto (s, t)$, for t fixed.) Then Fubini's theorem implies that $\{\lambda_t\}$ is a (T, μ) disintegration of λ. As in the case of Lebesgue measure on $\mathbb{R}^2$, the image measure $T\lambda$ is not sigma-finite unless ν is a finite measure. So it is handy that the definition of a (T, μ)-disintegration does not require μ to be the image measure $T\lambda$. Moreover, there is no way to get a disintegration with almost all λ_t probability measures if ν is not finite. □

One grudge held against disintegrations concerns existence. The abstract Kolmogorov approach to conditioning requires only pure measure theory; disintegrations, in general, are tainted by topological requirements, but they deliver more in terms of natural and useful properties. There is the usual trade-off: stronger requirements give stronger properties. We believe that the extra generality sacrificed by restricting to situations in which disintegrations exist will not be missed in many statistical applications.

We have found the following version of the existence theorem quite adequate, even though it is not the most general possible. We require that λ be a Radon measure (also known as a tight measure) on a metric space. That is, λ is a Borel measure for which $\lambda K < \infty$ for each compact K and $\lambda B = \sup_{K \subseteq B} \lambda K$, the supremum being taken over compact sets, for each Borel set B. For example, a finite Borel measure on a complete, separable metric space is Radon—see Theorem 1.4 of BILLINGSLEY (1968).

THEOREM 1. (EXISTENCE THEOREM) *Let λ be a sigma-finite Radon measure on a metric space $\mathcal{X}$ and let T be a measurable map from $\mathcal{X}$ into $(\mathcal{T}, \mathcal{B})$. Let μ be a sigma-finite measure on $\mathcal{B}$ that dominates the image measure $T\lambda$. If $\mathcal{B}$ is countably generated and contains all the singleton sets $\{t\}$, then λ has a (T, μ)-disintegration. The λ_t measures are uniquely determined up to an almost sure equivalence: if $\{\lambda_t^*\}$ is another (T, μ)-disintegration then $\mu\{t \in \mathcal{T} : \lambda_t \neq \lambda_t^*\} = 0$.*

Notice that the uniqueness assertion is much stronger than the almost sure uniqueness of $\mathbb{P}(X \mid T = t)$ for each integrable X in the Kolmogorov approach to conditioning. It requires existence of a single μ-negligible set N such that $\lambda_t A = \lambda_t^* A$ for all $t \notin N$ and *all* Borel sets A.

The proof of existence is just difficult enough to intimidate the typical graduate student, even though versions of it appear in many texts. We sketch a proof in the Appendix, to make the point that, with the possible exception of one topological/measure-theoretic fact, the argument is within the reach of most graduate probability courses.

On occasion, one works only with conditional probabilities of events involving another measurable map S into a space $(\mathcal{S},\mathcal{C})$. In such a case one needs the disintegrating measures defined only on the sigma-field $\mathcal{A}_0$ on $\mathcal{X}$ generated by the map $\Psi=(S,T)$ into $\mathcal{S}\otimes\mathcal{T}$. If the image measure $\Psi(\lambda)$ has a disintegration $\{\nu_t\}$ with respect to the coordinate projection onto $\mathcal{T}$ and the measure μ, and if the complement $\Psi(\mathcal{X})^c$ of the range of Ψ has zero outer $\Psi(\lambda)$ measure, then the disintegrating measures can be pulled back to $\mathcal{A}_0$ using the definition $\nu_t=\Psi(\lambda_t)$. Compare with Loève (1978, Section 30.2). It is easy to see that $\Psi(\mathcal{X})^c$ necessarily has zero inner measure. If it is not in the product sigma-field there might be some difficulty in arguing for zero outer measure. If λ were a Radon measure the set would have zero outer measure, but in that case why would one want to settle for less than the full disintegration for λ?

A few simple facts about disintegrations make them easy to work with. First let us be precise about when the disintegrating measures are probabilities. In essence, to get conditional probabilities one has only to standardize the disintegrating measures. The only subtlety is that standardization cannot work on a set of infinite or zero measure.

Theorem 2. *Let λ have a (T,μ)-disintegration $\{\lambda_t\}$, with λ and μ each sigma-finite.*

(i) *The image measure $T\lambda$ is absolutely continuous with respect to μ, with density $\lambda_t\mathcal{X}$.*

(ii) *The measures $\{\lambda_t\}$ are finite for μ-almost all t if and only if $T\lambda$ is sigma-finite.*

(iii) *The measures $\{\lambda_t\}$ are probabilities for μ-almost all t if and only if $\mu=T\lambda$.*

(iv) *If $T\lambda$ is sigma-finite then $(T\lambda)\{\lambda_t\mathcal{X}=0\}=0$ and $(T\lambda)\{\lambda_t\mathcal{X}=\infty\}=0$. For $T\lambda$-almost all t, the measures*

$$\tilde{\lambda}_t(\cdot)=\frac{\lambda_t(\cdot)}{\lambda_t\mathcal{X}}\{0<\lambda_t\mathcal{X}<\infty\}$$

are probabilities that give a T-disintegration of λ.

Proof. We abbreviate "for μ-almost all t" to "mod μ", and write $\ell(t)$ for the total mass, $\lambda_t\mathcal{X}$, of λ_t. For nonnegative measurable g,

$$(T\lambda)g=\lambda^x g(Tx)=\mu^t\lambda_t^x g(Tx)=\mu^t g(t)\ell(t) \tag{3}$$

As a service to readers who may still be getting used to our notation, we could write the last equalities as

$$\int g(t)(T\lambda)(dt)=\int g(Tx)\lambda(dx)=\int\int g(Tx)\lambda_t(dx)\mu(dt)=\int g(t)\ell(t)\mu(dt)$$

The simplification in the last equality occurs because $g(Tx) = g(t)$ for λ_t-almost all x, and that $g(t)$ can be brought outside the innermost integral as a constant—exactly what intuition says conditional distributions should allow.

For (i): If $g \geq 0$ and $\mu g = 0$ then $g(t) = 0 \bmod \mu$, whence $g(t)\ell(t) = 0 \bmod \mu$. In particular, every μ-negligible set is also $T\lambda$-negligible. Equation (3) is the formal statement that $\ell(t)$ is the density.

For (ii): Sigma-finiteness of a measure is equivalent to the existence of a strictly positive real-valued function with a finite integral. In particular, there exists an $h > 0$ for which $\mu h < \infty$. If $\ell(t) < \infty \bmod \mu$, the function $g(t) = h(t)/(1 + \ell(t))$ is strictly positive mod μ and $(T\lambda)g \leq \mu h < \infty$, which makes $T\lambda$ sigma-finite. Conversely, if $(T\lambda)k < \infty$ for some strictly positive k then $k(t)\ell(t) < \infty \bmod \mu$, by (i), which gives finiteness of $\ell(t) \bmod \mu$.

For (iii): If $\ell(t) = 1 \bmod \mu$ then equation (3) shows that $(T\lambda)g = \mu g$. By assumption, μ is always sigma-finite. For the converse, let h be strictly positive with $\mu h < \infty$ as in the previous paragraph. Choosing $g(t) = h(t)\{\ell(t) < 1\}$ in (3) and using the assumption that $T\lambda = \mu$ gives

$$\infty > \mu^t h(t)\{\ell(t) < 1\} = \mu^t h(t)\{\ell(t) < 1\}\ell(t)$$

which implies $\mu\{\ell < 1\} = 0$. A similar argument shows that $\mu\{\ell > 1\} = 0$.

For (iv): From (ii) we have $\mu\{l = \infty\} = 0$, so that (i) gives $(T\lambda)\{l = \infty\} = 0$. Take $g(t) = \{\ell(t) = 0\}$ in (3) to show that $(T\lambda)\{l = 0\} = 0$. For nonnegative measurable f, we then have

$$\begin{aligned}\lambda f &= \mu^t \lambda_t f \\ &= \mu^t \ell(t) \tilde{\lambda}_t f + \mu^t(\{\ell(t) = 0\}\lambda_t f) + \mu^t(\{\ell(t) = \infty\}\lambda_t f) \\ &= (T\lambda)^t \tilde{\lambda}_t f + 0 + 0\end{aligned}$$

The second term is zero because λ_t is the zero measure when $\ell(t) = 0$. The third term is zero because $\mu\{\ell = \infty\} = 0$. □

Caution! The result in part (i) can be most misleading when the image measure $T\lambda$ is not sigma-finite. For example, if T projects Lebesgue measure λ on $\mathbb{R}^2$ onto a coordinate axis, the image measure is not sigma-finite; it gives infinite measure to every set of nonzero one-dimensional Lebesgue measure. In one sense the function $\lambda_t \mathbb{R}^2 \equiv \infty$ is the correct Radon–Nikodym density, but the integration theory for such an extremely infinite measure is delicate and of little use; every set has image measure either zero or infinity. It would perhaps be better to insist that a density be finite almost everywhere, to avoid bad measures of this type. Only when $T\lambda$ is sigma-finite can it sensibly be used as the mixing measure μ. (The reader should exercise similar caution when interpreting part (ii) of the next Theorem.)

Notice that the construction for part (iv) can be applied more generally. If μ is dominated by a measure ν with a finite density $d\mu/d\nu = m(t)$, then λ has a (T, ν)-disintegration $\{\Lambda_t\}$ given by $\Lambda_t f = m(t)\lambda_t f$, because $\lambda f = \mu^t \lambda_t f = \nu^t(m(t)\lambda_t f)$.

Much of the convenience of working with disintegrations comes from the way they fit nicely with image measures and densities. Most of the following results are easy consequences of the special case treated by Theorem 2. We state them in full for future reference. See HOFFMANN-JØRGENSEN (1994, Section 10.11) for similar assertions for probability measures, proved in more traditional notation.

THEOREM 3. *Let λ have a (T,μ)-disintegration $\{\lambda_t\}$, and let ρ be absolutely continuous with respect to λ with a finite density $r(x)$, with each of λ, μ, and ρ sigma-finite.*

(i) *The measure ρ has a (T,m)-disintegration $\{\rho_t\}$ where each ρ_t is dominated by the corresponding λ_t, with density $r(x)$.*

(ii) *The image measure $T\rho$ is absolutely continuous with respect to μ, with density $\lambda_t r$.*

(iii) *The measures $\{\rho_t\}$ are finite for μ almost all t if and only if $T\rho$ is sigma-finite.*

(iv) *The measures $\{\rho_t\}$ are probabilities for μ almost all t if and only if $\mu = T\rho$.*

(v) *If $T\rho$ is sigma-finite then $(T\rho)\{\lambda_t r = 0\} = 0$ and $(T\rho)\{\lambda_t r = \infty\} = 0$. For $T\rho$-almost all t, the measures defined by*

$$\tilde{\rho}_t(f) = \frac{\lambda_t(fr)}{\lambda_t r}\{0 < \lambda_t r < \infty\} \tag{4}$$

are probabilities that give a T-disintegration of ρ.

PROOF. For (i) note that $\rho f = \lambda(rf) = \mu^t \lambda_t(rf)$. The other assertions follow from Theorem 2 via the equality $\rho_t \mathcal{X} = \lambda_t r$. □

The $\tilde{\rho}_t$ measures in part (v) are just the ρ_t of part (i) standardized to be probability measures, on the set $\{0 < \lambda_t r < \infty\}$ where standardization is possible. The complement of that set has zero $T\rho$ measure, so it wouldn't matter if we changed the definition of $\tilde{\rho}_t$ there. The disintegrating measures can be changed arbitrarily on a $T\rho$-negligible set without disturbing the disintegration.

The simple formula (4) is the general version of the familiar method for calculating conditional densities as a ratio of joint density to marginal density. It is more useful than the familiar formula because it does not require the conditioning variable to be a coordinate projection on a Euclidean space with Lebesgue measure playing the role of λ. LEHMANN (1959, Chapter 2, Lemma 6) used a special case of (4) in his treatment of exponential families.

EXAMPLE 3. Suppose P is a probability on $\mathbb{R}^k \otimes \mathbb{R}^{n-k}$ with density $p(x, y)$ with respect to Lebesgue measure. Disintegrate the dominating Lebesgue measure λ on $\mathbb{R}^n$ as in Example 2. Writing X for the projection onto $\mathbb{R}^k$, we have disintegrating probability measures P_x with (conditional) densities

$$p(x,y)/\lambda_x^y\, p(x,y) = \frac{p(x,y)}{\int p(x,y')dy'}$$

with respect to Lebesgue measure on $\mathbb{R}^k$, as taught in undergraduate classes.

Much theory in the statistical literature is based on this special case. One supposes that each member of a family $\{P_\theta : \theta \in \Theta\}$ is a probability on $\mathbb{R}^n$ and that T maps $\mathbb{R}^n$ smoothly into a lower-dimensional space $\mathbb{R}^k$. One assumes existence of another smooth map S from $\mathbb{R}^n$ into $\mathbb{R}^{n-k}$ such that $\phi(x) = (T(x), S(x))$ is smoothly invertible. Inferences on the family $\{P_\theta : \theta \in \Theta\}$ should then be equivalent to inferences on the family of image measures $\{\phi P_\theta : \theta \in \Theta\}$, for which there are densities with respect to Lebesgue measure on $\mathbb{R}^n$ in its role as the range space,

$$q(s,t,\theta) = p(\phi^{-1}(s,t),\theta)j(s,t) \tag{5}$$

Here $j(s,\ t)$ involves the Jacobian of the transformation ϕ. The conditioning variable is now one of the coordinate projections, for which the conditional density can be calculated as the ratio of joint to marginal densities,

$$\frac{q(s,t,\theta)}{\int q(s',t,\theta)ds'}$$

We have three qualms about this method for continuous distributions. First, it applies only to densities on Euclidean spaces. Second, it requires invention of an auxiliary map S that need be of no particular interest except that it builds the interesting T into a one-to-one transform of the data—one needs to force the conditioning variable to be a coordinate projection on a Euclidean space. Third, it requires extraneous smoothness assumptions about the conditioning map T, in order that the image measure might be absolutely continuous with respect to Lebesgue measure. As Theorem 3 shows, one needs none of these restrictive assumptions in order to derive a conditional density analogous to the ratio of joint to marginal densities. It is merely a matter of making a proper choice for the measure to use when calculating the "marginal" density. □

Many facts about abstract conditional expectations have analogs for disintegrations that make slightly stronger assertions under slightly more restrictive circumstances. We present just one example.

Conditional expectations given sigma-fields have the nesting property

$$\mathbb{P}(\mathbb{P}(X \mid \mathcal{F}_1) \mid \mathcal{F}_0) = \mathbb{P}(X \mid \mathcal{F}_0) \quad \text{when } \mathcal{F}_0 \subseteq \mathcal{F}_1$$

There is an analogous formula for disintegrations, which corresponds to the idea of taking conditional expectations over the variables that are discarded in pulling back to the coarser sigma-field.

EXAMPLE 4. Suppose λ is a sigma-finite measure on $(\mathcal{X}, \mathcal{A})$ with a $(T,\ \mu)$-disintegration $\{\lambda_t\}$, for a sigma-finite μ on $(\mathcal{T}, \mathcal{B})$, which in turn has a $(S,\ \nu)$-disintegration $\{\mu_s\}$ for a sigma-finite ν on $(\mathcal{S}, \mathcal{C})$. Here T is a measurable map from $\mathcal{X}$ into $\mathcal{T}$ and S is a measurable map from $\mathcal{T}$ into $\mathcal{S}$. Their composition $S \circ T$ is a measurable map from $\mathcal{X}$ into $\mathcal{S}$.

The measure λ has an $(S \circ T, \nu)$-disintegration $\{\gamma_s\}$ given symbolically by $\gamma_s = \mu_s^t \lambda_t$. One averages the λ_t disintegrations over all level sets that S maps onto s. That γ_s has the right averaging property follows from

$$\lambda f = \mu^t(\lambda_t f) = \nu^s(\mu_s^t(\lambda_t f)) \tag{6}$$

That it concentrates on the right level sets follows from the concentration properties of the other two disintegrations:

$$\begin{aligned}\nu^s \gamma_s^x\{S(Tx) \neq s\} &= \nu^s \mu_s^t(\{St = s\}\lambda_t^x(\{Tx = t\}\{S(Tx) \neq s\})) \\ &= \nu^s \mu_s^t \lambda_t^x\{St = s, Tx = t, S(Tx) \neq s\} \\ &= 0\end{aligned}$$

because the region of integration is an empty subset of $\mathcal{X} \otimes \mathcal{T} \otimes \mathcal{S}$.

Sigma-finiteness of λ implies existence of a strictly positive f for which λf is finite. Equality (6) then gives finiteness of $\mu_s^t \lambda_t f$ for ν-almost all s; the measure γ_s is sigma-finite for ν almost all s. □

3 Examples

In this section we present a small collection of examples that shows some of the benefits of treating conditioning as a matter of disintegration.

We start (Example 5) with the EM-algorithm, where it seems that one has to work explicitly with the conditional probability measure for a particular realization of a statistic. We set aside worries that the realization might fall in the negligible set where a meddling probabilist might decide to change the disintegrating measure. Once the conditional measure is fixed, the conditioning interpretation plays no further role in the analysis—our first example of conditioning has very little to do with conditioning.

We next turn to the Factorization Theorem for sufficient statistics (Example 6), a topic that first got us seriously interested in a more satisfactory way to work with conditioning. Most textbooks make it clear that the general version of the theorem is much too hard for general discussion. We feel the difficulty diminishes when one thinks of conditioning as disintegration.

The third example (Example 7) shows how the disintegrating measures can inherit invariance properties under a group of transformations.

The fourth example (Example 8) proves the converse of Basu's theorem about ancillary statistics. The proof is easy. It helped us to understand the need for something beyond independence from the sufficient statistic when we saw that the distributions concentrate on level sets defined by the disintegration.

The fifth example (Example 9) should be common knowledge to Bayesians, who know that posterior distributions are probability measures and not just collections of measurable functions that almost hang together in the right way. Their posteriors are disintegrating measures. To make life more interesting, we allow improper priors, with a reminder that even nonBayesians make use of Bayes estimators that guarantee

admissibility. One has only to be careful about infinite expectations at awkward moments.

The sixth example (Example 10) is an elementary Bayesian problem concerning the posterior distribution for a probability concentrated on two lines. We first present a non-rigorous, elementary method of solution, which we suspect would be the instinctive approach of most mathematical statisticians. (It was certainly how we initially solved the problem.) We then show how an even more general problem almost solves itself when properly framed: a small disappointment for anyone bent on demonstrating superiority of disintegrations, perhaps, but a genuine example of a method of solution that hadn't occurred to us before we started writing this paper. We recommend that our readers provide their own complete, rigorous solutions before looking at what we come up with.

Examples 11 and 12 come as a pair. They describe a marginalization paradox of STONE and DAWID (1972) that can afflict Bayesians with improper priors. We end up agreeing with Hartigan (1983, page 29), who pointed out the dangers in calculating marginal distributions by integration over unwanted variables. One must be careful when interpreting independence when probabilities are not finite.

In Example 13 we present a disintegration interpretation of the Gibbs sampler.

We could cite many stochastic process examples where the disintegration approach sheds light on complicated conditioning arguments. In an initial version of this paper we included one such application—the proof of continuity for the sample paths of martingales adapted to a Brownian filtration—and a referee pointed out other applications (interpretation of the strong Markov property; reflection principle for Brownian motion). For the sake of brevity, we decided to omit those examples from the final version, after realizing that stochastic process experts are unlikely to need further reminder of the advantages of working with regular conditional distributions or disintegrations.

EXAMPLE 5. The EM algorithm is often presented as a technique of maximum likelihood estimation for problems with missing data. For example, LITTLE and RUBIN (1987, page 127) describe it in the following way:

> Suppose as before that we have a model for the complete data Y, with associated density $f(Y \mid \theta)$ indexed by an unknown parameter θ. We write $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ where Y_{obs} represents the observed part of Y and Y_{mis} denotes the missing values. In this chapter we assume for simplicity that the data are [missing at random] and that the objective is to maximize the likelihood
>
> $$L(\theta \mid Y_{\text{obs}}) = \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) dY_{\text{mis}}$$
>
> with respect to θ.

The dY_{mis} here has presumably the symbolic meaning of whatever averaging is necessary to obtain the marginal density of Y_{obs}. The measure corresponding to

dY_{mis} would be Lebesgue measure if $f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta)$ were interpreted as a density with respect to Lebesgue measure on a product of Euclidean spaces. (A similar interpretation is needed for the $d\boldsymbol{x}$ at the top of page 96 of the WU (1983) paper.)

In situations where the observed data are given as some arbitrary function of Y, one must concoct a Y_{mis} so that the pair $(Y_{\text{obs}}, Y_{\text{mis}})$ becomes a one-to-one function of Y. The density for Y then transforms into a joint density for $(Y_{\text{obs}}, Y_{\text{mis}})$, in much the same way as in Example 3, and then the problem fits into a framework where conditioning can be handled by elementary means.

We would argue that in problems where data are naturally modelled as a function $T(x)$ on a probability space $(\mathcal{X}, \mathcal{A}, P_\theta)$ it is an unnecessary artifice to invent a missing function merely to accommodate EM theory to elementary methods of conditioning. One should instead start from a family of probability densities $\{p(x,\theta) : \theta \in \Theta\}$ with respect to a sigma-finite measure λ, which has a disintegration $\{\lambda_t\}$ with respect to (T, μ). The image measure has density

$$\phi(t,\theta) = \lambda_t^x p(x,\theta)$$

with respect to μ. For a given t, the maximum likelihood method seeks a θ to maximize $\phi(t,\theta)$.

More generally, one could consider the problem of maximizing a function

$$G(\theta) = \gamma^x g(x,\theta)$$

where $\{g(x,\theta) : \theta \in \Theta\}$ is a family of positive functions, and γ is a sigma-finite measure for which $0 < G(\theta) < \infty$ for each θ.

The generalized EM algorithm consists of repeated application of two steps that improve upon an initial guess θ_0 for the value maximizing G. Let Q_θ be the probability measure with density

$$q(x,\theta) = g(x,\theta)/G(\theta)$$

with respect to γ. In the E-step one calculates the expectation

$$L_0(\theta) = Q_{\theta_0}^x \log g(x,\theta)$$

In the M-step one maximizes L_0, or at least finds a θ_1 for which

$$L_0(\theta_1) > L_0(\theta_0)$$

The two steps are guaranteed to give

$$G(\theta_1) > G(\theta_0)$$

because

$$\begin{aligned} 0 &< L_0(\theta_1) - L_0(\theta_0) \\ &= Q_{\theta_0}^x \log\left(\frac{q(x,\theta_1)G(\theta_1)}{q(x,\theta_0)G(\theta_0)}\right) \\ &= \log\frac{G(\theta_1)}{G(\theta_0)} - \gamma^x q(x,\theta_0)\log\frac{q(x,\theta_0)}{q(x,\theta_1)} \end{aligned}$$

The last term is the Kullback–Leibler distance between Q_{θ_0} and Q_{θ_1}, which is positive by Jensen's inequality.

One then repeats the two steps, with θ_1 taking over the role of θ_0. And so on. □

The next example is the perfect illustration of how a disintegration proof can be built by analogy with simple arguments for a discrete case. The exposition is slightly more complicated than we would have liked, because we chose not to ignore some subtleties concerning division by zero. (The reader might care to ponder how these subtleties are usually taken care of in textbook proofs for the discrete case.) We have been told that a proof similar to ours appears in a book of Borovkov, but we have not yet been able to find that book.

EXAMPLE 6. The intuitive definition of sufficiency says that a statistic T is sufficient for a family of probability measures $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ if the conditional distributions given T do not depend on θ. We avoid technical "difficulties concerning the behavior of conditional probabilities" by interpreting the definition to mean existence of a shared disintegration $\{P_t\}$. That is, $P_t(\cdot)$ should serve as a conditional distribution $\mathbb{P}_\theta(\cdot \mid T = t)$ for every θ.

Most often one checks for sufficiency by means of a factorization criterion, whose general proof has a forbidding reputation. LEHMANN (1959) approached the proof in a most sensible manner, by discussing first the discrete version, then a special case of the continuous version (using methods based on the transformation idea described in Example 3), and finally presenting (Section 2.6) the full-blown Radon–Nikodym approach of HALMOS and SAVAGE (1949) only after careful measure-theoretic preparation. We were struck by the differences between the proofs in the various cases. The disintegration interpretation allows us to use the same idea for all cases.

Consider first the proof that factorization implies sufficiency in the discrete case. Here each $\mathbb{P}_\theta$ is defined on a finite set $\mathcal{X}$ with probabilities that factorize as

$$p(x, \theta) = \mathbb{P}_\theta\{x\} = g(Tx, \theta)h(x)$$

for some statistic T. The conditional expectations are then obtained as simple ratios. For fixed t,

$$\mathbb{P}_\theta(f \mid T = t) = \frac{\Sigma_{Tx=t} g(Tx, \theta)h(x)f(x)}{\Sigma_{Tx=t} g(Tx, \theta)h(x)} = \frac{\Sigma_{Tx=t} h(x)f(x)}{\Sigma_{Tx=t} h(x)}$$

The factors involving g in the numerator and denominator have cancelled out, leaving a ratio that does not depend on θ. (Might there be any problem with 0/0 here?)

The last formula has a simple disintegration interpretation. Let us regard $p(x, \theta)$ as the density of $\mathbb{P}_\theta$ with respect to counting measure λ on $\mathcal{X}$. With μ as counting measure on $\mathcal{T}$, the (T, μ)-disintegration of λ has λ_t equal to counting measure on $\{T = t\}$. The last displayed ratio is just the expectation

$$P_t(f) = \frac{\lambda_t(fh)}{\lambda_t h}\{0 < \lambda_t h < \infty\}$$

We have included the explicit indicator function to avoid one 0/0 problem. The upper bound on $\lambda_t h$ is automatic for the case of a finite set $\mathcal{X}$, but is needed already for countably infinite sets.

Now consider the general case, where $\mathbb{P}_\theta$ has density $g(Tx,\theta)h(x)$ with respect to a general sigma-finite measure λ. Suppose λ has a (T,μ)-disintegration $\{\lambda_t\}$ for some sigma-finite μ. For a fixed θ, Theorem 3 shows that $\mathbb{P}_\theta$ has a T-disintegration $\{\mathbb{P}_{\theta,t}\}$, where

$$\mathbb{P}_{\theta,t}(f) = \frac{\lambda_t^x g(Tx,\theta)h(x)f(x)}{\lambda_t^x g(Tx,\theta)h(x)}\{0 < \lambda_t^x g(Tx,\theta)h(x) < \infty\}$$

By parts (ii) and (v) of the same Theorem and the concentration property of the $\{\lambda_t\}$,

$$0 < \lambda_t^x g(Tx,\theta)h(x) = g(t,\theta)\lambda_t h < \infty \text{ for } T\mathbb{P}_\theta\text{-almost all } t$$

We are therefore almost everywhere justified in cancelling out a $g(t, \theta)$ factor from numerator and denominator to get

$$\mathbb{P}_{\theta,t}(\cdot) = P_t(\cdot) \text{ for } T\mathbb{P}_\theta\text{-almost all } t$$

where $P_t(\cdot)$ is defined by (7), just as in the discrete case except for the changed meaning of λ_t. The disintegration property is unaffected if we change the disintegrating measures for a $T\mathbb{P}_\theta$-negligible set of t. The $\{P_t\}$ also define a T-disintegration for $\mathbb{P}_\theta$, as required for sufficiency.

For the converse it is useful to replace λ by a dominating probability measure of the form $\mathbb{P} = \Sigma_i 2^{-i}\mathbb{P}_{\theta_i}$, for some countable subfamily $\{\mathbb{P}_{\theta_i}\}$ of $\mathcal{P}$. (A device due to HALMOS and SAVAGE, 1949—see Theorem 2 in the Appendix of LEHMANN, 1959). What matters is that the common disintegrating probabilities $\{P_t\}$ for each $\mathbb{P}_\theta$ also provide a T-disintegration for $\mathbb{P}$, because

$$(T\mathbb{P})^t P_t f = \sum_i 2^{-i}(T\mathbb{P}_{\theta_i})^t P_t f = \mathbb{P}f$$

If we write $g(t, \theta)$ for the density of $T\mathbb{P}_\theta$ with respect to $T\mathbb{P}$, we have

$$\begin{aligned}\mathbb{P}_\theta f &= (T\mathbb{P}_\theta)^t P_t f && \text{definition of } \{P_t\}\\ &= (T\mathbb{P})^t g(t,\theta)P_t f && \text{definition of } g(t,\theta)\\ &= \mathbb{P}^x g(Tx,\theta)f(x)\end{aligned}$$

the last equality holding because $\{P_t\}$ is also a disintegration for $\mathbb{P}$. Thus $\mathbb{P}_\theta$ has density $g(Tx,\theta)$ with respect to $\mathbb{P}$, and density $g(Tx,\theta)d\mathbb{P}/d\lambda$ with respect to λ. □

Sometimes the disintegration can be identified by an appeal to symmetry, or to an invariance argument, with the uniqueness of disintegrations simplifying the formal proof.

EXAMPLE 7. Let P be a probability measure on a space $\mathcal{X}$. Suppose a probability measure P is invariant under a group $\mathcal{G}$ of transformations on $\mathcal{X}$. That is, $gP = P$ for all g in $\mathcal{G}$. Suppose also that the sets $\{T = t\}$ are invariant under $\mathcal{G}$. Does it follow that the conditional distributions are also invariant under $\mathcal{G}$?

For example, the standard normal distribution on $\mathbb{R}^2$ is invariant under rotations about the origin. The statistic $T(x_1, x_2) = x_1^2 + x_2^2$ is constant on circles centered at the origin, sets that are invariant under rotations. The conditional distributions are uniform around the circles, a fact that it usually demonstrated by means of a calculation with Jacobians. In higher dimensions the argument becomes quite messy. How much easier it would be if we could deduce the form the conditional distributions directly from invariance considerations.

The argument succeeds if $\mathcal{G}$ can be replaced by a countable subclass $\mathcal{G}_0$. Suppose measures invariant under $\mathcal{G}_0$ are necessarily also invariant under the whole of $\mathcal{G}$. Then if P is invariant under $\mathcal{G}$, the conditional distributions P_t must also be invariant under $\mathcal{G}$, for TP-almost all t. The proof depends on the uniqueness of disintegrations.

For each bounded measurable f on $\mathcal{X}$, and each g in $\mathcal{G}$,

$$\begin{aligned} Pf &= (gP)f && \text{invariance of } P \\ &= P(f \circ g) && \text{definition of image measure } gP \\ &= (TP)^t P_t(f \circ g) && \text{disintegration} \\ &= (TP)^t (gP_t) f && \text{definition of image measure } gP_t \end{aligned}$$

When P_t concentrates on the set $\{T = t\}$, so does gP_t. It follows that $\{gP_t : t \in \mathcal{T}\}$ is another disintegration for P. By uniqueness of disintegrations, there exists a TP-negligible set $\mathcal{N}_g$ such that $gP_t = P_t$ for all t in $\mathcal{N}_g^c$. Cast out a sequence of negligible sets—one for each member of $\mathcal{G}_0$—to deduce that, for TP-almost all t, the probability measure P_t is invariant under $\mathcal{G}_0$, and hence invariant under $\mathcal{G}$. □

The Borel paradox (KOLMOGOROV, 1993, page 50) is the classic example of an unjustifiable appeal to invariance for the construction of conditional distributions. POLLARD (1996, Chapter 5) has explained the source of the difficulty, using the language of disintegration.

EXAMPLE 8. Suppose T is sufficient for $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ on Ω, with disintegrating measures $\{P_t\}$. Let f and g be bounded real functions on $\mathbb{R}$, and let S be another statistic. Define $G(t) = P_t g(S)$. Then

$$\mathbb{P}_\theta f(T) g(S) = (T\mathbb{P}_\theta)^t f(t) P_t g(S) = \mathbb{P}_\theta f(T) G(T) \tag{8}$$

In particular, $g(S)$ and $G(T)$ have the same $\mathbb{P}_\theta$ expectation, $C(\theta)$.

If S is ancillary, then the expected value $C(\theta)$ is equal to a constant C. If T is boundedly complete, then the assertion $\mathbb{P}_\theta G(T) = C$ implies that $\mathbb{P}_\theta$ concentrates on $\{G(T) = C\}$ for each θ, whence

$$\mathbb{P}_\theta f(T) G(T) = \mathbb{P}_\theta f(T) \cdot C = \mathbb{P}_\theta f(T) \mathbb{P}_\theta g(S)$$

It follows that S is independent of T. That is the BASU (1955, 1958) theorem.

Conversely, if S is independent of T under each $\mathbb{P}_\theta$, then by choosing $f = G$ in (8) we get

$$(\mathbb{P}_\theta G(T))^2 = \mathbb{P}_\theta G(T)^2$$

so that $\mathbb{P}_\theta$ concentrates on the level set $\Omega(\theta) = \{G(T) = C(\theta)\}$. If there were θ_0 and θ_1 for which $C(\theta_0) \neq C(\theta_1)$, we would have a partition of Θ into two nonempty subfamilies,

$$\Theta_0 = \{\theta : \mathbb{P}_\theta \text{ concentrated on } \Omega(\theta_0)\}$$
$$\Theta_1 = \{\theta : \mathbb{P}_\theta \text{ concentrated on } \Omega(\theta_0)^c\}$$

with the corresponding families of probability measures supported by disjoint subsets of Ω. If such a partition of Θ is assumed impossible then $C(\theta)$ must be a constant, that is, S must be ancillary. That is the converse to Basu's theorem. □

Basu's results can be proved without the use of disintegrations. For us, the advantage of the proof with disintegrations is clean definition of the two sets Θ_0 and Θ_1.

Bayesians work with conditional distributions, by choice. Decision theorists often apply Bayesian arguments. The next Example broadens slightly the scope of a venerable admissibility argument, as used by EATON (1992), for example, by removing unnecessary sigma-finiteness assumptions. Our approach is based on an idea explained to us by John Hartigan.

EXAMPLE 9. Let $\{P_t : t \in \mathcal{T}\}$ be a family of probability measures on $\mathcal{X}$. If the map $t \mapsto P_t f$ is measurable for nonnegative measurable f on $\mathcal{X}$, and if π is a probability (a prior distribution) on $\mathcal{T}$, then a probability measure $\mathbb{Q}$ can be defined on $\mathcal{X} \otimes \mathcal{T}$ by

$$\mathbb{Q}g = \pi^t P_t^x g(x, t)$$

The coordinate maps X and T have joint distribution $\mathbb{Q}$. The $\{P_t\}$ have the interpretation of a T-disintegration of $\mathbb{Q}$, that is, the conditional distribution of X given $T = t$ is P_t. The X-disintegration of $\mathbb{Q}$ defines the Bayesian posterior distribution $\mathbb{Q}_x(\cdot) = \mathbb{Q}(\cdot \mid X = x)$.

If each P_t has a density $p(x, t)$ with respect to a sigma-finite μ on $\mathcal{X}$, then

$$\mathbb{Q}g = \pi^t \mu^x p(x, t) g(x, t)$$

That is, $\mathbb{Q}$ has density $p(x, t)$ with respect to the product measure $\mu \otimes \pi$. The product measure has the trivial disintegration

$$(\mu \otimes \pi)_x = \pi,$$

if we abuse notation as in Example 2. It follows from Example 3 that $\mathbb{Q}_x$ has density

$$p(x, t)/\pi^t p(x, t) \tag{9}$$

with respect to π.

Given a nonnegative loss function $L(t, d)$ on $\mathcal{T} \otimes \mathcal{T}$, a Bayes estimator $\delta(x)$ can be defined by the value of α that minimizes the posterior expected loss $\mathbb{Q}_x^t L(t, \alpha)$,

$$\mathbb{Q}_x^t L(t, \delta(x)) = \inf_{\alpha} \mathbb{Q}_x^t L(t, \alpha) \quad \text{for each } x \tag{10}$$

Even nonBayesians are interested in such estimators because they enjoy a number of nice decision theoretic properties. For example, suppose δ has finite Bayes risk

$$\mathbb{Q}L(t, \delta(x)) = \pi^t P_t^x L(t, \delta(x)) = \nu^x \mathbb{Q}_x^t L(t, \delta(x)) < \infty$$

where ν stands for the marginal distribution of X under $\mathbb{Q}$. Suppose also that $\delta^*(x)$ is another estimator with smaller expected loss,

$$P_t^x L(t, \delta^*(x)) \leq P_t^x L(t, \delta(x)) < \infty \quad \text{for } \pi\text{-almost all } t \tag{11}$$

Then strict inequality can hold only on a π-negligible set, for otherwise

$$\begin{aligned} 0 &> \pi^t P_t^x (L(t, \delta^*(x)) - L(t, \delta(x))) \\ &= \nu^x \mathbb{Q}_x^t (L(t, \delta^*(x)) - L(t, \delta(x))) \end{aligned} \tag{12}$$

The defining property of $\delta(x)$ requires the last integrand to be everywhere nonnegative, which gives a contradiction to the inequality (12).

The preceding argument has little to do with π or any of the disintegrating $\{\mathbb{Q}_x\}$ being probability measures, nor with ν being the marginal X distribution. It is valid for any (X, ν) disintegration and any sigma-finite π (an improper prior), provided the Bayes estimator defined by equality (10) has finite Bayes risk $\mathbb{Q}L(t, \delta(x))$.

For example, if P_t is the Bin(n, t) distribution and π is the improper prior with density $t^{-1}(1-t)^{-1}$ with respect to Lebesgue measure λ on $(0, 1)$, then $\mathbb{Q}$ has density

$$p(x, t) = \binom{n}{x} t^{x-1}(1-t)^{n-x-1}$$

with respect to $\lambda \otimes \mu$, where μ is counting measure on $\{0, 1, \ldots, n\}$. The marginal measure $\nu = X\mathbb{Q}$ is not sigma-finite; it puts infinite mass at 0 and at n. Nevertheless, $\mathbb{Q}$ has an (X, μ)-disintegration with $\mathbb{Q}_x$ having density $p(x, t)$ with respect to λ. Notice that $\mathbb{Q}_x$ is a finite measure for $1 \leq x \leq n-1$, and both $\mathbb{Q}_0$ and $\mathbb{Q}_n$ are infinite (but sigma-finite).

Let $L(t, d) = (t-d)^2$. For $1 \leq x \leq n-1$ the usual argument shows that the estimator $\delta(x) = x/n$ minimizes the posterior expected loss. It also minimizes

$$\mathbb{Q}_0^t L(t, \alpha) = \int_0^1 (t-\alpha)^2 t^{-1}(1-t)^{n-1}\, dt$$

for the trivial reason that the integral is finite only when α equals zero. Similar trivial

reasoning applies to $\mathbb{Q}_n$. The estimator δ is Bayes for the improper prior π, with a finite Bayes risk,

$$\sum_{x=0}^{n} \int_0^1 \binom{n}{x} t^{x-1}(1-t)^{n-x-1}(t-x/n)^2\,dt < \infty$$

The inequality corresponding to (11) could hold only on a π-negligible set. As both sides of that inequality would be polynomials in t, the negligible set would have to be empty: a contradiction. The Bayes estimator is admissible for quadratic loss. □

EXAMPLE 10. Suppose a distribution P on $\mathbb{R}^2$ concentrates on two straight lines, L_1 and L_2, neither of them orthogonal to the x-axis. Suppose the total mass p_i that P assigns to L_i is distributed according to a density g_i with respect to Lebesgue measure along the line. An observation (X, Y) is taken from P giving a point with $X = x_0$. What is the conditional probability that the point lies on the line L_1?

The elementary method approximates $\{X = x_0\}$ by $\{x_0 \le X \le x_0 + \delta\}$, for a small positive δ, then argues that

$$\begin{aligned} P((X, Y) \in L_1 \mid X = x_0) &\approx \frac{P((X, Y) \in L_1, x_0 \le X \le x_0 + \delta)}{P(x_0 \le X \le x_0 + \delta)} \\ &\approx \frac{p_1\alpha_1 g_1(x_0)\delta}{p_1\alpha_1 g_1(x_0)\delta + p_2\alpha_2 g_2(x_0)\delta} \end{aligned}$$

where $1/\alpha_i$ is the absolute value of the cosine of the angle between L_i and the x-axis. The small δ factors cancel out, leaving an equality

$$P((X, Y) \in L_1 \mid X = x_0) = \frac{p_1\alpha_1 g_1(x_0)}{p_1\alpha_1 g_1(x_0) + p_2\alpha_2 g_2(x_0)}$$

in the limit.

Write $H_1(x_0)$ for the last ratio. If one wants a totally rigorous derivation using the Kolmogorov approach, one can easily check the defining property analogous to (1),

$$\mathbb{P}\{(X, Y) \in L_1\}\{X \in B\} = \mathbb{P}H_1(X)\{X \in B\}$$

for all Borel sets B. Alternatively, one might appeal to some sort of abstract differentiation theorem to guarantee existence of the limiting ratio and justify its interpretation as a conditional probability.

Both rigorous derivations would obscure the simple form of the conditional probability distribution $\mathbb{P}\{\cdot \mid X = x_0\}$, which puts mass $H_1(x_0)$ at the point where L_1 intersects the line where $X = x_0$, and mass $1 - H_1(x_0)$ at the corresponding point on L_2. Provided (X, Y) does not land at the intersection of the two lines, this conditional probability distribution gives the asserted mass to line L_1. We prefer an argument that identifies the conditional distribution directly, rather than have it emerge indirectly from a calculation of uncertain rigor.

To determine the X-disintegration of P we need first to be precise about what we mean by a distribution with a density with respect to Lebesgue measure along a line in $\mathbb{R}^2$. Bring everything back to the x-axis $\mathcal{X}$, by regarding Lebesgue measure along L_i as the image of α_i times Lebesgue measure μ along $\mathcal{X}$. The geometry of the lines enters only through the α_i factors. The measure with density $p_i g_i(\cdot)$ with respect to Lebesgue measure on L_i is just the image of the measure μ_i with density $h_i(x) = \alpha_i p_i g_i(x)$ with respect to μ. Now we can forget all about lines and Lebesgue measure, and solve a more general problem.

Suppose μ is a sigma-finite measure on $\mathcal{X}$ and that $h_1, \ldots, h_k$ are nonnegative integrable functions on $\mathcal{X}$. Let $\psi_1, \ldots, \psi_k$ be measurable maps into another space $\mathcal{Y}$. Let μ_i be the finite measure with density h_i with respect to μ. Let Q_i be the image of μ_i under the map ϕ_i that takes x onto $(x, \psi_i(x))$. That is, Q_i is the result of sliding μ_i to live on the graph of ψ_i in the product space $\mathcal{X} \otimes \mathcal{Y}$. Define P to be the sum of the Q_i. What is the conditional distribution $P(\cdot \mid X = x)$?

Formally,

$$Pf = \sum_{i=1}^{k} \phi_i(\mu_i)(f) = \sum_{i=1}^{k} \mu^x h_i(x) f(x, \psi_i(x)) \tag{13}$$

Taking the μ outside the last sum we immediately get a representation of P as an (X, μ)-disintegration,

$$Pf = \mu^x P_x f$$

where P_x is the measure that puts mass $h_i(x)$ at the points $(x, \psi_i(x))$ for $i = 1, \ldots, k$. Notice that P_x is not a probability, but it does live on the set $\{X = x\}$. To make the disintegrating measures probabilities, we need to standardize as prescribed by Theorem 2. For $i = 1, \ldots, k$ the conditional probability measure $P(\cdot \mid X = x)$ (that is, the X-disintegrating measure) puts mass $h_i(x)/(h_1(x) + \cdots + h_k(x))$, except at the negligible set of x values where the denominator is zero. □

The result from the previous Example is a solution to a Bayesian problem posed to us by John Hartigan. Le Cam (1986, page 477) has used an analogous disintegration to establish a bound on Hellinger affinities for convex hulls. With reference to this result, Donoho and Liu (1991, page 644) remarked that "Le Cam has established a fact which seems, at first, quite similar to ... but is in fact far deeper". The case of finite convex combinations is a simple consequence of an identity like (13); the general case is a consequence of a general disintegration.

For a measure λ on a product space $\mathcal{X} \otimes \mathcal{Y}$ it is traditional to use the name $\mathcal{X}$-marginal for the image of λ under the map X that projects onto the $\mathcal{X}$ coordinate space. If λ happens to be a product of probability measures, $P \otimes \nu$, the $\mathcal{X}$-marginal equals P. One can safely refer to both $P\{x \in \mathcal{X} : x \in A\}$ and $(P \otimes \nu)\{(x, y) \in \mathcal{X} \otimes \mathcal{Y} : x \in A\}$ as "the probability that X lies in the set A". However, if ν is not a probability measure, the $\mathcal{X}$-marginal of λ does not equal P. At worst, ν might not

even be a finite measure, in which case the image measure assigns mass ∞ to every A with $PA > 0$. In this situation there is real danger in thinking of the $\mathcal{X}$- and $\mathcal{Y}$-coordinates as being independent, or even in thinking of P as the distribution of X. Bayesians with a penchant for improper priors should be particularly aware of this problem.

EXAMPLE 11. Suppose (X, Y) has strictly positive probability density $f(x, y)$ with respect to Lebesgue measure on the unit square $(0,1) \otimes (0,1)$. Then, in traditional notation, X has marginal density $f_X(x) = \int_0^1 f(x,y)\,dy$ and the conditional distribution of Y given $X = x$ has conditional density $f_{Y|X}(y \mid x) = f(x,y)/f_X(x)$. Given $X = x$ and $Y = y$, let the Z distribution be the constant multiple $1/f(x, y)$ times Lebesgue measure on $\mathbb{R}$. The joint (improper) distribution of (X, Y, Z) is equal to three-dimensional Lebesgue measure λ on $(0,1) \otimes (0,1) \otimes \mathbb{R}$.

With λ expressed as a product of Lebesgue measure on each coordinate space, we might be tempted to think of X, Y, and Z as independent, each uniformly distributed. Indeed, for a product of proper probability measures, the coordinate maps are independent random variables with those probabilities as their marginal distributions. However, our Example involves products of improper distributions: the Z marginal is Lebesgue measure—the improper uniform distribution on the real line—and conditional on $Z = z$, the pair (X, Y) is uniformly distributed on the square. (That is, we have a Z-disintegration of λ with Lebesgue measure on the unit square as disintegrating measure.) Since the last conditional distribution does not depend on z, we might conclude that (X, Y) is uniform on the square, so that X and Y are independent and uniform on (0, 1). Or should we use the (X, Y)-marginal measure, which is very infinite, as the joint distribution? Or should we stick with the original $f(x,y)$ density?

Is there any paradox in (X, Y) appearing to have several different joint distributions? We think not.

The confusion arises because Lebesgue measure on $(0,1) \otimes (0,1) \otimes \mathbb{R}$ can be disintegrated in many different ways. The (X, Y)-image measure is not sigma-finite; it cannot be used as the mixing measure in an (X, Y)-disintegration. With the image measure no longer a candidate, there are many equally plausible mixing measures and disintegrations, giving many different plausible answers for the joint distribution of X and Y and for the conditional distribution of Y given X.

When one works only with probability measures, all arguments lead back to the same joint (marginal) distributions. With infinite measures, different derivations can lead to different measures. One should exercise some care in bestowing the title of joint distribution. □

The rather obvious sort of distinction in the last Example can become much more puzzling when buried within more complicated collections of marginal and conditional distributions, as in the following *marginalization paradox* of STONE and DAWID (1972). When their constructions are expressed as explicit assertions about disintegrations, the flaw behind the paradox is quickly revealed.

EXAMPLE 12. Consider a measure defined on $(\mathbb{R}^+)^4$, specified in the traditional way by means of distributions of random variables as coordinate projections. Let

$$\begin{aligned} \Phi &\sim L = \text{Lebesgue measure on } \mathbb{R}^+, \\ (\Theta \mid \Phi = \phi) &\sim \text{probability density } \pi(\theta) \text{ with respect to } L, \\ (X \mid \Theta = \theta, \Phi = \phi) &\sim \text{exponential, with mean } 1/(\theta\phi), \\ (Z \mid X = x, \Theta = \theta, \Phi = \phi) &\sim \text{exponential, with mean } 1/(\phi x). \end{aligned}$$

The random vector (Φ, Θ, X, Z) has a joint (improper) distribution λ with density

$$f(\phi, \theta, x, z) = \pi(\theta)\theta\phi^2 x e^{-(\theta+z)\phi x} \tag{14}$$

That is, we have defined a sigma-finite measure on $(\mathbb{R}^+)^4$ with density f with respect to the product L^4 of Lebesgue measures on the coordinate spaces.

The (Φ, Θ, Z)-marginal distribution is sigma-finite with density

$$f(\phi, \theta, z) = \frac{\theta\pi(\theta)}{(\theta+z)^2}$$

and the (Φ, Z) marginal is sigma-finite with density

$$f(\phi, z) = I(z) = L^\theta\left(\frac{\theta\pi(\theta)}{(\theta+z)^2}\right)$$

Notice that neither density depends on ϕ; both marginal measures are products having the measure L on the Φ-axis as a factor.

Disintegration with respect to the (Φ, Z) marginal measure gives the $(\Theta \mid \Phi, Z)$ conditional probability density

$$f(\theta \mid \phi, z) = \frac{\theta\pi(\theta)}{(\theta+z)^2 I(z)} \tag{15}$$

Noting that the last conditional distribution does not depend on ϕ, we might be tempted to conclude that the $(\Theta \mid Z)$ conditional density is

$$f(\theta \mid z) \stackrel{?}{=} \frac{\theta\pi(\theta)}{(\theta+z)^2 I(z)} \tag{16}$$

Example 4 would justify such a conclusion if the (Φ, Z)-marginal had a disintegration with the Z-marginal as mixing measure. Unfortunately the Z-marginal is not sigma-finite. Assertion (16) is based on a false analogy with the result for averaging over disintegrating *probability* measures.

Something has gone wrong already, but now let us repeat the same sort of reasoning along a different path. The (Θ, X, Z)-marginal is sigma-finite with density

$$f(\theta, x, z) = \frac{2\theta\pi(\theta)}{x^2(\theta+z)^3}$$

and the (X, Z)-marginal is sigma-finite with density $2J(z)/x^2$, where

$$J(z) = L^\theta\left(\frac{\theta\pi(\theta)}{(\theta+z)^3}\right)$$

The marginal distribution of (Θ, X, Z) has a (X, Z)-disintegration with the $(\Theta \mid X, Z)$ conditional probability density

$$f(\theta \mid x, z) = \frac{\theta\pi(\theta)}{(\theta+z)^3 J(z)} \tag{17}$$

Again we might be tempted to interpret the lack of dependence on one variable, x, as meaning that the $(\Theta \mid Z)$ conditional density is

$$f(\theta \mid z) \stackrel{?}{=} \frac{\theta\pi(\theta)}{(\theta+z)^3 J(z)} \tag{18}$$

And again we would be misled by the analogy with Example 4. The Z-marginal is still not sigma-finite, no matter how it is calculated.

Formulae (16) and (18) appear contradictory—a paradox. We would explain the paradox by pointing out that neither formula represents a disintegration of the (Θ, Z)-marginal with the Z-marginal as mixing measure. There is no such disintegration. The assertions (15) and (17) are fine; each statement gives a disintegration with respect to a sigma-finite measure. In both cases, the trouble comes when we are tempted to throw away one of the conditioning variables, leaving just the variable Z, whose image measure is not sigma-finite. We must live with the fact that distributions conditional on Z are not determined uniquely. So there is no such thing as *the* $(\Theta \mid Z)$ conditional density. Indeed, in this case, much as in Example 11, the (Θ, Z) image measure is not sigma-finite. So in constructing a (Θ, Z)-disintegration for the joint distribution of (Φ, Θ, X, Z), we are left with a rather arbitrary choice of what measure to use as the (Θ, Z)-mixing measure. If we then regard our choice of (Θ, Z)-mixing measures as the "joint distribution" for that pair of variables, then clearly we can arrive at many different "conditional distributions" for $(\Theta \mid Z)$.

More concisely, by integrating out variables in different orders we have constructed two distinct sequences of disintegrations for calculating $L \otimes L \otimes L \otimes L(fg)$:

$$L^\phi L^z\left[I(z)L^\theta\left[\frac{\theta\pi(\theta)}{(\theta+z)^2 I(z)}L^x[x(\theta+z)^2\phi^2 e^{-(\theta+z)\phi x}g(\phi,\theta,z,x)]\right]\right]$$
$$= L^x\left[\frac{1}{x^2}L^z\left[2J(z)L^\theta\left[\frac{\theta\pi(\theta)}{(\theta+z)^3 J(z)}L^\phi\left[\frac{\phi^2}{2}(\theta+z)^3 x^3 e^{-(\theta+z)x\phi}g(\phi,\theta,z,x)\right]\right]\right]\right]$$

All integrals correspond to probability measures, except for the first "$L^\phi(\cdots)$" and the second "$L^x(\cdots)$" and "$L^z(\cdots)$". If $L^\theta(\pi(\theta)/\theta)$ were finite, then we could also have

standardized J to be a probability density. It is futile to try to interpret the probability measures as *the* conditional distributions. □

EXAMPLE 13. Terms like "Markov chain Monte Carlo" and "Markov sampling" refer to methods for generating random samples from given distributions by running Markov chains. Although such methods have quite a long history, they have become the subject of renewed interest in the last decade, particularly with the introduction of the "Gibbs sampler" by GEMAN and GEMAN (1984), who used the method in a Bayesian approach to image reconstruction. The Gibbs sampler itself has enjoyed a recent surge of intense interest within statistics community, spurred by GELFAND and SMITH (1990), who applied the Gibbs sampler to a wide variety of inference problems.

Recall that a distribution P being *stationary* for a Markov chain $X_0, X_1, \ldots$ means that, if $X_0 \sim P$, then $X_n \sim P$ for all n. The theoretical foundation of Markov sampling methods is the convergence in distribution of a Markov chain to its stationary distribution: If a Markov chain $X_0, X_1, \ldots$ has stationary distribution P, then under quite general conditions (involving irreducibility and aperiodicity), the distribution of X_n for large n is close to P. Thus, in order to generate an observation from a desired distribution P on $\mathcal{X}$, we find a Markov chain $X_0, X_1, \ldots$ on $\mathcal{X}$ that has P as its stationary distribution. The theory then suggests that running or simulating the chain until a large time n will produce a random variable X_n whose distribution is close to the desired P. By taking n large enough, in principle we obtain a value that may for practical purposes be considered a random draw from the distribution P.

The Gibbs sampler is a way of constructing a Markov chain having a desired stationary distribution. To illustrate the idea, consider a product space $\mathcal{X} = \mathcal{S} \otimes \mathcal{T}$ with coordinate maps S and T. The problem is to generate an observation from a given probability measure P on $\mathcal{X}$. We assume that both S- and T-disintegrations of P exist, giving conditional probability distributions that we will denote as $P(\cdot \mid S = \cdot)$ and $P(\cdot \mid T = \cdot)$. To perform a Gibbs sampler, start with any initial point (S_0, T_0). Then generate S_1 from the conditional distribution $P(\cdot \mid T = T_0)$, and generate T_1 from the conditional distribution $P(\cdot \mid S = S_1)$. Continue on in this way, generating S_2 from the conditional distribution $P(\cdot \mid T = T_1)$ and T_2 from the conditional distribution $P(\cdot \mid S = S_2)$, and so on. Then the distribution P is stationary for the Markov chain $\{(S_n, T_n) : n = 0, 1, \ldots\}$. To see this, suppose $(S_0, T_0) \sim P$. In particular, T_0 is distributed according to the T-marginal of P, so that, since S_1 is drawn from the conditional distribution of S given $T = T_0$, we have $(S_1, T_0) \sim P$. Now we use the same reasoning again: S_1 is distributed according to the S-marginal SP, so that $(S_1, T_1) \sim P$.

Here is a general formulation of the Gibbs sampler in terms of disintegrations. Suppose we wish to simulate an observation from a probability measure P on a space $\mathcal{X}$. The Gibbs sampler consists of a sequence of "moves" that tell us how to choose a new point X_{n+1}, given a current point X_n. For each map T for which a T-disintegration $\{P_t\}$ of P exists, there is a corresponding "T-move", which is

performed as follows: Given the current point $X_n \in \mathcal{X}$, draw the next point X_{n+1} according to the distribution $P_{T(X_n)}$. A T-move leaves the measure P invariant, that is,

$$P^x P_{T(x)} f = Pf$$

In fact, this is just a restatement of the averaging property required in the definition of disintegration: defining $g(t) = P_t f$, we have

$$Pf = (TP)^t P_t f = (TP)^t g(t) = P^x g(T(x)) = P^x P_{T(x)} f$$

Thus, for any map T, the Markov chain $X_0, X_1, \ldots$ produced by a succession of T-moves has the desired distribution P as a stationary distribution. However, such a chain would stay on the same level set of the map T forever, that is, we would have $X_n \in \{x : T(x) = T(X_0)\}$ for all n. To have convergence in distribution to P starting from an arbitrary initial distribution, we must perform moves using more than one disintegration. That is the Gibbs sampler: given a sequence of disintegrations, the Gibbs sampler is a performance of the corresponding moves. For example, given two maps S and T, we could alternate making S-moves and T-moves. Or we could flip a coin at each iteration to decide whether to make an S-move or a T-move. There is no need to restrict to product spaces and coordinate maps as in the illustrative simple setting above. □

4 Other notions of conditioning

We hope we have convinced you that the existence of a disintegration is very convenient in many statistical problems. However we do not wish to give the impression that we never feel the need to condition on sigma-fields. After all, the expectation of an integrable X with respect to the disintegrating $\mathbb{P}_t$ is just a version of the abstract Kolmogorov conditional expectation. More precisely, if we define $G(t) = \mathbb{P}_t X$ then, at least for bounded measurable functions H,

$$\begin{aligned} \mathbb{P}H(T)X &= (T\mathbb{P})^t \mathbb{P}_t(H(T)X) \\ &= (T\mathbb{P})^t H(t)\mathbb{P}_t X \quad \text{because } \mathbb{P}_t \text{ concentrates on } \{T = t\} \\ &= \mathbb{P}H(T)G(T) \end{aligned} \tag{19}$$

That is, $Y = G(T)$ is the (almost surely) unique random variable that is measurable with respect to the sigma-field $\mathcal{G}$ generated by T for which

$$\mathbb{P}WX = \mathbb{P}WY \quad \text{all bounded } \mathcal{G}\text{-measurable } W \tag{20}$$

As the proof in the Appendix shows, the fact that $\mathbb{P}_t$ concentrates on $\{T = t\}$ is actually equivalent to equality (19) for a suitable countable collection of H functions. No topology is needed there. It is in the interpretation of $\mathbb{P}(\cdot \mid \mathcal{G})$ as an expectation with respect to a probability measure that topology intervenes, as a way of sorting out problems with uncountable collections of negligible sets. If we are concerned with the

conditional expectations of only countably many random variables—as in the theory of discrete-time martingales, for example—then there is no need to bring in topological tools to manage the almost sure equivalences. However, in statistical problems many surprises can lie hidden in the formulations using conditioning on sigma-fields.

Consider, for example, the concept of sufficiency. One could call a sub-sigma-field $\mathcal{G}$ sufficient for a family $\mathcal{P}$ of probability measures on $(\Omega, \mathcal{F})$ if, for each bounded $\mathcal{F}$-measurable random variable X there exists a single $\mathcal{G}$-measurable Y for which equality (20) holds for every $\mathbb{P}$ in $\mathcal{P}$. That is the standard rigorous definition. As BURKHOLDER (1961) showed, the definition allows some disturbing consequences, such as the possibility of a sufficient $\mathcal{G}$ being contained in a finer sigma-field $\mathcal{G}^*$ that is not sufficient for $\mathcal{P}$. If the intuition behind sufficiency says that $\mathcal{G}$ contains all the information about which $\mathbb{P}$ in $\mathcal{P}$ we are sampling from, then how can $\mathcal{G}^*$ be telling us something extra? Apparently, the abstract definition has let a few nonintuitive beasts through the gate. In the case of a dominated family no such problem can exist—compare with the factorization theorem of Example 6.

In some situations even the abstract definition is too concrete; the interpretation of $Y = \mathbb{P}(X \mid \mathcal{G})$ as a random variable (or as an equivalence class of random variables) becomes superfluous. We can identify Y with a transition operator γ, mapping $L^1(\mathbb{P}, \mathcal{F})$ into $L^1(\mathbb{P}, \mathcal{G})$, identified by the analog of equality (20),

$$\langle \gamma X, W \rangle = \langle X, W \rangle \quad \text{all } W \text{ in } L^\infty(\mathbb{P}, \mathcal{G})$$

And then we can dispense with Y altogether and express conditioning properties purely in terms of a transition operator. DAWID (1980) chose something similar as the best way to deal with the general form of conditional independence.

Finally, one can dispense with the interpretation of the domains of probability measures as families of random variables on a specific Ω set, and treat conditioning as a transition map between abstract spaces, as in HARTIGAN's (1983) development of Bayes theory, or LE CAM's (1986) theory for convergence of experiments.

By stripping away assumptions unnecessary for the development of a particular statistical or probabilistic idea, one gains in generality and sometimes even in insight. We would claim that disintegrations offer more insight into something like the factorization criterion for sufficiency than a collection of more elementary calculations for specific cases, sometimes involving unnecessary technical assumptions to accommodate the details of a particular method. In the same way, disintegrations could be regarded as overly restrictive, involving unnecessary topological assumptions in many abstract conditioning arguments using Radon–Nikodym derivatives. And so on.

Conditioning is one of the most important ideas of probability and statistics. It is needed at many different levels of understanding. We see great value in there also being many ways of formalizing its mathematical description, each suited to a different purpose.

5 History

The concepts of conditioning have a long history, which we cannot claim to have researched carefully. The best we can do is offer some references that might help those who wish to pursue the topic further.

LOÈVE (1978, Section 30.2) mentioned that the problem of existence of regular conditional probabilities was "investigated principally by Doob", but he cited no specific reference. DOOB (1953, page 624) cited a counterexample to the unrestricted existence, which also appears in the exercises to Section 48 of the 1969 printing of HALMOS (1950). Doob's remarks suggest that the original edition of the Halmos book contained a slightly weaker form of the counterexample. Doob also noted that the counterexample destroyed a claim made in (DOOB, 1938), an error pointed out by Dieudonné (no citation) and Andersen and Jessen (no citation)—perhaps in their (1946) paper?

BLACKWELL (1956) cited DIEUDONNÉ (1948) as the source of a counterexample for unrestricted existence of a regular conditional probabilities. Blackwell also proved existence of regular conditional distributions for (what are now known as) Blackwell spaces. The proof given by DELLACHERIE and MEYER (1978) uses the same sort of regularity properties on the underlying space.

HOFFMANN-JØRGENSEN (1994, page 162) asserted that KOLMOGOROV (1933) was the first to establish existence of regular conditional distributions (for "ordinary random vectors"). We could not find this result in Kolmogorov's book; indeed he stressed (page 50) that the conditional probability $P_u(B)$ was determined only up to an almost sure equivalence. Chapter 10 of the Hoffman-Jørgensen book contains an exposition of the best disintegration theorem available, a result due to PACHL (1978). Pachl cited a number of earlier papers on disintegrations.

PARTHASARATHY (1967, Sections V.7 and V.8) cited notes of Varadarajan for his existence proof for a disintegration.

A mention of the names Doob and Kuratowski by WILLIAMS (1979, page 100) was drawn to our attention by a referee, but we were unable to trace further—Williams cited no works of those two authors. Probably DOOB (1953) was intended, but we can only guess about Kuratowski. (Maybe the Topology book?)

The key idea in all proofs of existence of regular conditional distributions is that of compact approximation—existence of a class of approximating sets with properties analogous to the class of compact sets in a metric space—as a means for deducing countably additivity from finite additivity. PFANZAGL and PIERLO (1969) developed a systematic theory of compact approximation. They were cited in the *Note Historique* by BOURBAKI (1969), who also gave credit to Ryll-Nardzewski for disintegration (no citation), perhaps in some point-process context. In point process theory disintegrations appear as Palm distributions—conditional distributions given a point of the process at a particular position (KALLENBERG, 1969).

PFANZAGL (1979) gave a condition under which a regular conditional distribution can be obtained by means of the elementary "limit of ratio of probabilities".

The BARNDORFF-NIELSEN, BLAESILD and ERIKSEN (1989) book contains much material on the invariance properties of conditional distributions, which we have not yet studied in detail.

6 Appendix: Existence of disintegrations

Here is a condensed proof of the Existence Theorem 1, based on ideas from DELLACHERIE and MEYER (1978, page 78). We agree with them that "The theorem on disintegration of measures has a bad reputation, and probabilists often try to avoid the use of conditional distributions ... But it really is simple and easy to prove."

The assumptions let us reduce the proofs of both existence and uniqueness to the case where $\mathcal{X}$ is compact and both λ and μ are finite measures. (Partition $\mathcal{T}$ into countably many disjoint sets B_i, each of finite μ measure. Partition each $T^{-1}B_i$ into sets $N_i, K_{i1}, K_{i2}, \ldots$, with $\lambda N_i = 0$ and each K_{ij} compact. For existence, construct finite disintegrating measures for the restriction of λ to K_{ij} and μ restricted to B_i, then piece together the restrictions. Notice that each disintegrating measure will be sigma-finite, being constructed from countably many finite measures concentrated on disjoint sets. For uniqueness, combine the trivial result for the restriction of λ to a negligible set with the result for compact sets.)

Define a finite measure ν (the image of λ under the map that takes x onto (x, Tx)) on $\mathcal{A} \otimes \mathcal{B}$ by $\nu h(x,t) = \lambda h(x, Tx)$. It lives on the graph of T, in the sense that

$$\nu\{(x,t) : Tx \neq t\} = 0 \tag{21}$$

This assertion follows from the countable generation property, and the fact that $\mathcal{B}$ contains all the singleton sets. (Let $\mathcal{B}_0$ be the countable subclass that generates $\mathcal{B}$. For each $t \in \mathcal{T}$, the singleton $\{t\}$ is equal to $\{t\} = \cap\{B \in \mathcal{B}_0 : t \in B\}$, which implies that $\{(x,t) : Tx \neq t\} = \cup_{B \in \mathcal{B}_0}\{Tx \notin B, t \in B\}$. For fixed B in $\mathcal{B}_0$,

$$\nu\{Tx \in B, t \notin B\} = \lambda\{Tx \in B, Tx \notin B\} = 0$$

The set $\{Tx \neq t\}$ is a countable union of ν-negligible sets.)

Now we use compactness to avoid the problem, mentioned at the start of Section 2, with uncountable families of negligible sets. The trick is to reduce countable additivity to a condition involving only countably many assertions about conditional expectations. On the real line one can determine measures from the values taken by their distribution functions at a countable dense set. On more general spaces, the following consequence of the Riesz representation theorem, and of the fact that there exists a sequence of functions dense in the space of all continuous real functions on $\mathcal{X}$ (under the uniform metric), suffices.

If $\mathcal{X}$ is a compact metric space then there exists a countable family $\mathcal{C}_0$ of non-negative, continuous functions on $\mathcal{X}$ such that

(i) $\mathcal{C}_0$ is closed under addition

(ii) for each additive functional $\ell : \mathcal{C}_0 \to \mathbb{R}^+$ there exists a unique Borel measure L such that $\ell(f) = Lf$ for each f in $\mathcal{C}_0$.

For fixed f in $\mathcal{C}_0$, the map $g \mapsto \nu(f(x)g(t))$ defines a measure on $\mathcal{B}$, which is dominated by μ because $|\nu(f(x)g(t))| \le C\lambda|g(Tx)| = C(T\lambda)|g|$ for some constant C that bounds f. Write $\lambda_t f$ for a density of this measure with respect to μ:

$$\nu(f(x)g(t)) = \mu^t(g(t)\lambda_t f)$$

(As a function of t, the $\lambda_t f$ integral corresponds to the Kolmogorov conditional expectation of f.) For almost all t, the map $f \mapsto \lambda_t f$ is nonnegative and additive, and hence corresponds to a measure on $\mathcal{A}$. Invoke a generating-class argument to deduce that $t \mapsto \lambda_t^x h(x,t)$ is measurable, for bounded measurable h, and $\nu h = \mu^t \lambda_t^x h(x,t)$. In particular, $\lambda f = \mu^t \lambda_t^x f(x)$ for each bounded, $\mathcal{A}$-measurable f. Put $h(x,t) = \{(x,t) : Tx \ne t\}$ to deduce from property (21) that $\mu^t \lambda_t\{x : Tx \ne t\} = 0$. Consequently, $\lambda_t\{x : Tx \ne t\} = 0$ for μ almost all t.

For uniqueness, suppose we have two disintegrations, $\{\lambda_t\}$ and $\{\lambda_t^*\}$, of a finite Radon measure λ on a compact metric space. Consider an f in $\mathcal{C}_0$. Define $B_f = \{t \in \mathcal{T} : \lambda_t f < \lambda_t^* f\}$. The two disintegrations of λ give

$$\mu^t\{t \in B_f\}\lambda_t f = \lambda\{T \in B_f\}f = \mu^t\{t \in B_f\}\lambda_t^* f$$

Deduce that $\mu B_f = 0$, that is, $\lambda_t f \ge \lambda_t^* f$ for almost all t. Argue analogously to get the reverse inequality. Cast out countably many negligible sets as f ranges over $\mathcal{C}_0$, to deduce that λ_t and λ_t^* can be different measures only for a μ-negligible set of t values.

7 Acknowledgements

We thank John Hartigan for suggesting several troublesome examples. We are also grateful for the comments of two careful referees and the editor, which helped us in our final revision.

References

ANDERSEN, E. S. and B. JESSEN (1946), Some limit theorems on integrals in an abstract set, *Det Kongelige Danske Videnskabernes Selskab, Matematisk-Fysiske Meddelelser*, Bind 22, no. 14.

BARNDORFF-NIELSEN, O. E., P. BLAESILD and P. S. ERIKSEN (1989), *Decomposition and invariance of measures, and statistical transformation models*, Vol. 58 of *Springer Lecture Notes in Statistics*, Springer-Verlag, New York.

BASU, D. (1955), On statistics independent of a complete sufficient statistic, *Sankhyā* **15**, 377–380.

BASU, D. (1958), On statistics independent of a sufficient statistic, *Sankhyā* **20**, 223–226.

BLACKWELL, D. (1956), On a class of probability spaces, in: J. NEYMAN (ed.) *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, University of California Press, Berkeley, pp. 1–6.

BOURBAKI, N. (1969), *Intégration*, Vol. IX of *Éléments de mathématique*, Hermann, Paris. (Fascicule XXXV).

DELLACHERIE, C. and P. A. MEYER (1978), *Probabilities and potential*, North-Holland, Amsterdam.

DIEUDONNÉ, J. (1948), Sur le théorème de Lebesgue-Nikodym, III, *Annales de l'Institut Fourier (Grenoble)* **23**, 25–53.
DONOHO, D. L. and R. C. LIU (1991), Geometrizing rates of convergence, II, *Annals of Statistics* **19**, 633–667.
DOOB, J. L. (1938), Stochastic processes with integral-valued parameter, *Transactions of the American Mathematical Society* **44**, 87–150.
DOOB, J. L. (1953), *Stochastic processes*, Wiley, New York.
EATON, M. (1992), A statistical diptych: admissible inferences-recurrence of symmetric Markov chains, *Annals of Statistics* **20**, 1147–1179.
GELFAND, A. E. and A. F. M. SMITH (1990), Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.
GEMAN, S. and D. GEMAN (1984), Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.
HALMOS, P. R. (1950), *Measure theory*, Van Nostrand, New York, NY. (July 1969 reprinting).
HALMOS, P. R. and L. J. SAVAGE (1949), Application of the Radon-Nikodym theorem to the theory of sufficient statistics, *Annals of Mathematical Statistics* **20**, 225–241.
HARTIGAN, J. A. (1983), *Bayes theory*, Springer, New York.
HOFFMANN-JØRGENSEN, J. (1994), *Probability with a view toward statistics*, Vol. 2, Chapman and Hall, New York.
KALLENBERG, O. (1969), *Random measures*, Akademie-Verlag, Berlin. (US publisher: Academic Press).
KOLMOGOROV, A. N. (1933), *Foundations of probability*, Chelsea, New York, NY. Second English Edition 1950.
LE CAM, L. (1986), *Asymptotic methods in statistical decision theory*, Springer, New York.
LEHMANN, E. L. (1959), *Testing statistical hypotheses*, Wiley, New York. Later edition published by Chapman and Hall.
LITTLE, R. J. A. and D. B. RUBIN (1987), *Statistical analysis with missing data*, Wiley, New York.
LOÈVE (1978), *Probability theory*, Springer, New York. Fourth Edition, Part II.
PACHL, J. (1978), Disintegration and compact measures, *Mathematica Scandinavica* **43**, 157–168.
PARTHASARATHY, K. R. (1967), *Probability measures on metric spaces*, Academic, New York.
PFANZAGL, J. (1979), Conditional distributions as derivatives, *Annals of Probability* **7**, 1046–1050.
PFANZAGL, J. and N. PIERLO (1969), *Compact systems of sets*, Vol. 16 of *Springer Lecture Notes in Mathematics*, Springer-Verlag, New York.
POLLARD, D. (1996), *Probability explained* (Unpublished book manuscript).
STONE, M. and A. P. DAWID (1972), Un-Bayesian implications of improper Bayes inference in routine statistical problems, *Biometrika* **59**, 369–375.
TJUR, T. (1974), *Conditional probability distributions*, Vol. 2 of *Lecture Notes*, Institute of Mathematical Statistics, University of Copenhagen.
WILLIAMS, D. (1979), *Diffusions, Markov processes, and martingales*, Vol. 1, Wiley, New York.
WINTER, B. B. (1979), An alternate development of conditioning, *Statistica Neerlandica* **33**, 197–212.
WU, C. F. J. (1983), On the convergence properties of the EM algorithm, *Annals of Statistics* **10**, 95–103.

Received: September 1994. Revised: April 1996.