Some reflections on and experiences with SPLIFs

Heinrich v. Weizsäcker

Abstract

Starting from the uniqueness question for mixtures of distributions this review centers around the question under which formally weaker assumptions one can prove the existence of SPLIFs, in other words perfect statistics and tests. We mention a couple of positive and negative results which complement the basic contribution of David Blackwell in 1980. Typically the answers depend on the choice of the set theoretic axioms and on the particular concepts of measurability.

The following pages describe some of my personal experiences and motivations connected to the subject of David Blackwell's 1980 note 'There are no Borel SPLIFs' [2]. I hope to show how this two page paper with a mysterious title (SPLIF stands for 'strong probability limit identification function') leads us directly to the foundations of the probabilistic formalism.
The measure theoretic language of probability provided by S. Ulam and N. Kolmogorov is used by many without much attention. We all use English without being experts in grammar. But for every language there always are and always should be those who study meticulously the rules and the scope of what could be expressed using the framework given by these rules. In the case of the measure theoretic language this is part of what I always was interested in. Blackwell's paper touches in an extremely elegant way the bounds of this framework.
Given this interest, why study measures on a space of measures? Of course a statistician trained in using Kolmogorov's framework first thinks (with or without some distrust) of Thomas Bayes' dictum *By chance I mean the same as probability* ([1], p.376) , when he refers to the problem of finding 'the chance that a probability lies between two given bounds'. For me the motivation came from a slightly different angle, namely from the theorem of

de Finetti or rather from the effort to understand this and similar extremal integral representation results from a more abstract point of view.

Let $(\Theta, \mathcal{T})$ be a parameter set with a $\sigma-$field $\mathcal{T}$, let $\{p_\vartheta\}_{\vartheta\in\Theta}$ be a family of probability measures on the measurable space $(\Omega, \mathcal{B})$ such that $\vartheta \longmapsto p_\vartheta(B)$ is measurable for every $B \in \mathcal{B}$. For the sake of simplicity of the exposition we shall make the regularity assumption that $(\Theta, \mathcal{T})$ and $(\Omega, \mathcal{B})$ are Borel subsets of Polish spaces with the induced Borel structure.

The mixtures

$$p_\lambda = \int_\Theta p_\vartheta \, \lambda(d\vartheta) \tag{1}$$

where λ varies over all probability measures on Θ form a convex set H. Under the above assumption every extreme point of H is of the form p_ϑ. This follows e.g. from prop. 1 a) in the survey [22]. Conversely in most concrete cases no p_ϑ is a mixture of the others and thus

$$ex\ H = \{p_\vartheta\}_{\vartheta\in\Theta}. \tag{2}$$

(An interesting exception is given by the mixtures of Weibull distributions: A Weibull distribution with shape parameter p is a mixture of Weibull distributions with shape parameter q as long as $p < q$, cf. [10], p. 480.) Under the condition (2) the mixing equation (1) is a Choquet type representation, i.e. the representation of a general element of H as a mixture of the extremal elements.

Naturally we ask the question: Is the representation (1) unique? In general the answer is negative. Examples of nonuniqueness are given by the set of one-dimensional centered distributions or more generally sets of martingale distributions. As a trivial special case consider $p_1 = \frac{1}{3}\delta_{-2} + \frac{2}{3}\delta_1, p_2 = \frac{2}{3}\delta_{-1} + \frac{1}{3}\delta_2, p_3 = \frac{1}{2}(\delta_{-1} + \delta_1), p_4 = \frac{1}{2}(\delta_{-2} + \delta_2)$. Then $\frac{1}{2}(p_1 + p_2) = \frac{2}{3}p_3 + \frac{1}{3}p_4$. Clearly the uniqueness means that the map $\lambda \longmapsto p_\lambda$ induces an affine isomorphism of the space $Prob(\Theta)$ of probability measures on Θ onto H, or equivalently that the cone $\mathbb{R}_+H$ is isomorphic to the cone $\mathcal{M}_+(\Theta)$ of positive bounded measures on Θ. How can such an isomorphism arise? Looking around I realized that a really remarkable situation turns out to be quite frequent: Some probabilistic limit theorem gives a function $\varphi : \Omega \longrightarrow \Theta$ such that

$$p_\vartheta\{\omega : \varphi(\omega) = \vartheta\} = 1 \tag{3}$$

for every $\vartheta \in \Theta$. Then the unique representing measure of an element q of H is given by $\lambda = q \circ \varphi^{-1}$. These functions φ are precisely the SPLIFs, or perfect statistics.

In de Finetti's case one has the strong law of large numbers, or the decreasing martingale theorem, in more general ergodic decompositions one has Birkhoff's ergodic theorem, in the case of Gibbs measures the decreasing

martingale theorem which corresponds to the thermodynamic limit. Dynkin [6] gave a systematic study of such situations. In his language the SPLIF is given by a 'H-sufficient' statistic. An interesting case in which the representation (1) is unique but a function φ with (3) does not necessarily exist arises in point process theory: p_ϑ is the law of the Poisson process given by the intensity measures ϑ. The mixtures correspond to Cox or compound Poisson processes (cf. Krickeberg [12]).
So we are (as I was) led to the question: What else is needed besides uniqueness in (1) in order to ensure the existence of a SPLIF? The direction of our search leads also to the concept of a PLIF, a 'probability limit identification function'. For the motivation let us start with an application of a SPLIF.

A remarkably general application of the existence of a SPLIF is given by one of the early successes of martingale theory: Doob's [5] consistency result for the posterior distributions: Let the $\sigma-$field $\mathcal{B}$ be generated by the union sub-$\sigma-$fields $\mathcal{B}_n$. Let λ be a prior and suppose that there is a $\lambda-$a.s. $\mathcal{B}-$measurable probability identifying function, i.e. a Borel map $\varphi : \Omega \longrightarrow \Theta$ such that

$$\lambda\{\vartheta \in \Theta : (3) \text{ holds}\} = 1. \tag{4}$$

Then the posterior probabilities $\hat{\lambda}_n(T|\mathcal{B}_n)$ converge to $1_T(\vartheta)$ $p_\vartheta-$a.s. for $\lambda-$a.s. $\vartheta \in \Theta$ and every measurable set $T \subset \Theta$. Since the topology on Θ has a countable base one gets

$$\lambda\{\vartheta : p_\vartheta\{\omega : \hat{\lambda}_n(\omega) \longrightarrow \varepsilon_\vartheta \text{ weakly}\} = 1\} = 1 \tag{5}$$

where ε_ϑ is the point mass in ϑ.
The intriguing fact is that for this consistency argument of Doob the probability identification (3) needs to work only for all ϑ outside a $\lambda-$nullset! Thus the function φ may be allowed to depend on λ. The existence of such a φ for each λ follows already from the existence of a PLIF, i.e. from asymptotic consistency in probability: Let d denote the metric on Θ. Suppose that there is a PLIF, i.e. there is a sequence (φ_n) of $\mathcal{B}_n-$measurable function $\varphi_n : \Omega \longrightarrow \Theta$ such that for every $\varepsilon > 0$

$$\lim_{n\to\infty} p_\vartheta\{\omega : d(\varphi_n(\omega), \vartheta) > \varepsilon\} = 0 \tag{6}$$

for all ϑ. Then given any prior λ on Θ it is easy to extract a subsequence such that

$$\lambda\{\vartheta : p_\vartheta\{\omega : d(\varphi_{n_k}(\omega), \vartheta) \longrightarrow 0\} = 1\} = 1$$

which implies the existence of a φ which satisfies (4).
But as we mentioned in the known concrete situations one gets even more: a SPLIF which does not involve any prior. Let us summarize:

Theorem 1 *Let $\vartheta \mapsto p_\vartheta$ be a transition kernel. Each of the following statements implies the next:*
(α) There is a Borel SPLIF, i.e. a Borel function φ which satisfies (3) for all $\vartheta \in \Theta$.
(β) There is a Borel PLIF, i.e. there is a sequence (φ_n) which satisfies (6) for all $\vartheta \in \Theta$.
(γ) For every prior λ there is a Borel function φ which satisfies (3) for λ-almost all $\vartheta \in \Theta$.

We are led to the

Question 1: *Is (α) implied by (β) or even by (γ) ?*

In order to understand the question better let us look at condition (γ) a little more closely. It can be reformulated in the following alternative way which led us in [14] to say that the family $\{p_\vartheta\}$ is 'orthogonality preserving', whereas a kernel with (α) was called 'completely orthogonal'.

($\gamma_\perp$) *For two orthogonal priors $\lambda_0 \perp \lambda_1$ the mixtures p_{λ_0} and p_{λ_1} are orthogonal as well.*

Another way to look at this is from the point of view of vector lattices: ($\gamma_\perp$) says that the mixing map $\lambda \mapsto p_\lambda$ is not only injective but it is also a lattice homomorphism from $\mathcal{M}_+(\Theta)$ into $\mathcal{M}_+(\Omega)$ where we recall that the lattice operations in the space of positive bounded measures can be defined via the Radon-Nikodym theorem

$$\frac{d\min(\mu,\nu)}{d\mu+\nu} = \min\Big(\frac{d\mu}{d\mu+\nu}, \frac{d\nu}{d\mu+\nu}\Big)$$

and similarly for the max.
Thus the property (γ) gives both from the Bayesian point of view (consistency of posterior distributions) and from the point of view of the vector lattice structure of the mixtures a quite natural way of the identifiability condition. The attractive feature of the condition (γ) resp. ($\gamma_\perp$) is that there are no limit theorems visible. But still there is the connection to consistency.
Here is a more precise reformulation of the consistency aspect of condition (γ). We mentioned that (γ) is implied by the existence of a sequence (φ_n) which is consistent in probability in the sense of (6). There is an interesting partial converse. David Preiss had the idea to use the concept of filters of countable type which I believe is due to Grimeisen [9] and Katetov [11]. Simply put, this class of filters can be characterized by being the smallest class of filters such that the *liminf* of a sequence of *liminf*-s along filters of countable type is again a *liminf* along a filter of countable type. Convergence

along such filters shares with convergence of sequences many properties like the dominated convergence theorem. In [14] (Theorem 4.1) it was shown that (γ) is equivalent to

(γ_c) *There is a family* $(\varphi_i)_{i\in\mathbb{N}}$ *of Borel maps from* Ω *to* Θ *and a filter* $\mathcal{F}$ *of countable type on* $\mathbb{N}$ *such that for every* $\vartheta \in \Theta$

$$\lim_{\mathcal{F}} p_\vartheta\{d(\varphi_i(\omega), \vartheta) > \varepsilon\} = 0\,. \tag{5'}$$

Note the fact that in (γ_c) no prior on Θ is involved !
So the assumption (γ) is fairly close to the existence of PLIFs; the difference being that in (γ_c) we take limits over a filter of countable type rather than the usual limit of a sequnce (which is the limit over the filter of cofinite sets in $\mathbb{N}$). Now let us try to reshape the condition (α), i.e. the existence of SPLIFs. A natural observation is that (α) is equivalent to the following condition (α_s).

(α_s) *If* Θ_0, Θ_1 *are two disjoint measurable subsets of* Θ *then there are disjoint measurable subsets* B_0, B_1 *of* Ω *such that* $p_{\vartheta_i}(B_i) = 1$ *whenever* $\vartheta_i \in \Theta_i$ *and* $i \in \{0, 1\}$.

In (α_s) the two sets $H_i = \{p_\lambda : \lambda \in Prob(\Theta_i)\}$ for $i = 0, 1$ are closed under mixtures and every element in H_0 is orthogonal to every element of H_1 under assumption (α_s). Thus $(\gamma) \Longrightarrow (\alpha)$ would hold provided the following question had a positive answer.

Question 2: *Let* $H_0, H_1 \subset Prob(\Omega, \mathcal{B})$ *be closed under mixtures and Borel sets for the topology of convergence in law. Suppose that every element of* H_0 *is orthogonal to every element of* H_1. *Do there exist disjoint sets* $B_0, B_1 \in \mathcal{B}$ *such that every element of* H_i *is concentrated on* B_i *for* $i = 0, 1$?

The following nice positive result was discovered independently by many authors: (e.g. Goullet de Rugy [7], Graf-Mägerl [8], Ornstein-Weiss [16], and [14]).

Theorem 2 *If, in question 2, the sets* H_i *are compact in the topology of convergence in law then the answer is "yes".*

The proof of Graf-Mägerl uses Choquet capacities, the proof in [14] mimicked a classical statistical minimax argument and gave even a quantitative version of the result. This gives a positive answer to question 1 under the additional assumption that the family $\{p_\vartheta\}$ is σ-compact in the topology of convergence in law.
Also, using the technique of filters of countable type we gave a positive answer to question 2 if one of the sets H_i is a singleton. This method implies

that under assumption (γ) the family has the following property (δ) which is the measurably parametrized version of a concept introduced by Dorothy Maharam-Stone [13] under the name 'uniform orthogonality'.

(δ) *There is a product measurable set* $B \subset \Omega \times \Theta$ *such that for every* $\vartheta \in \Theta$ *the section* $B_\vartheta \subset \Omega$ *satisfies* $p_\vartheta(B_\vartheta) = 1$ *and* $p_{\vartheta'}(B_\vartheta) = 0$ *for all* $\vartheta' \neq \vartheta$.

In 'statistical terms' this means that for every simple null-hypothesis there is a test of power 1. If Θ is the set of all lines in the plane Ω and each p_ϑ is a normal law concentrated on ϑ then $\{p_\vartheta\}$ satisfies (δ) but not (γ).
Coming back to question 2, another straightforward application of the same method yields a result of G. Mokobodzki [19]. It assumes the existence of a medial limit $m : [0,1]^{\mathbb{N}} \to [0,1]$, i.e. a universally measurable map m such that $\liminf_i z_i \leq m(z) \leq \limsup_i z_i$ for all $z = (z_i) \in [0,1]^{\mathbb{N}}$ which is measure affine: $m(\int z d\mu) = \int m(z) d\mu$ for every Borel probability measure on $[0,1]^{\mathbb{N}}$. It is known that medial limits exist under the continuum hypothesis and even under the weaker 'Martin's axiom'. Under the assumption of a medial limit Mokobodzki proved that question 2 has a positive answer if the sets H_i are analytic and the B_i are allowed to be universally measurable. Similarly under this assumption property (γ) implies the existence of a universally measurable SPLIF. This contains the older result of J. Štepan [20].
But is all this set theory necessary? Perhaps condition (γ) implies the existence of a Borel SPLIF after all? We are asking whether something can be proven within the standard framework of probability (Borel functions and no special axioms). The turning point was given by the following counter-example of David Blackwell [2]. The proof was a beautiful application of Baire category. The main idea had various applications and interpretations (cf. [4],[3]). A subset of a topological space has the property of Baire, if there is an open set $U \subset \Omega$ such that $B \Delta U$ is of first category.

Theorem 3 *(Blackwell) Let* $\Omega = \{0,1\}^{\mathbb{N}}$ *and consider for* $i = 0, 1$ *the set*

$$H_i = \{p \in Prob(\Omega) : p\{\omega_k = 1\} \underset{k \to \infty}{\longrightarrow} i\}.$$

Let $B \subset \Omega$ *such that* $p(B) = i$ *for all* $p \in H_i$ *for* $i = 0, 1$. *Then* B *cannot have the property of Baire. In particular* B *is not Borel.*

The σ-algebra of sets with the property of Baire is very large. In fact S. Shelah [17], improving a famous result of R. Solovay [18], proved that it is consistent with Zermelo-Fraenkel set theory (of course without the axiom of choice) that every subset of a Polish space has the property of Baire! As was

remarked already by Solovay in such a world many surprising things happen, like that the Banach space $L^1(\mu)$ is reflexive for every finite measure μ. Combining Blackwell's theorem with Mokobodzki's result we can add that no medial limits exist there.

Thus the answer to question 2 was negative. In an obvious sense question 2 is the analogue of question 1 for tests. This left the question 1, i.e. the case of perfect statistics still open. But David Preiss who had found an alternative proof of the Borel part of Blackwell's theorem improved his technique with the help of a very clever application of the (hard) measurable selection theorem for Borel sets with compact sections to prove the existence of a counterexample to question 1. When I talked to David Blackwell about this he stressed that if an example exists it should be possible to have an explicit description. This remark kept working in me and finally Dan Mauldin and I [15] managed to find a surprisingly simple strategy to directly construct for $\Theta = [0,1]$ uncountably many nonisomorphic families with property with a PLIF but without a SPLIF, by modifying the classical Bernoulli family on $\{0,1\}^{\mathbb{N}}$. On the other hand in [14] it was shown that every two families of diffuse measures with the same parameter set which allow a SPLIF can be transformed into each other by a Borel isomorphism of the observation spaces.

Let us come to an end asking two somewhat technical questions which are left open by the above discussion, and mentioning one final positive result:

Question 3: *Can one prove in Zermelo-Fraenkel set theory with the axiom of choice that the existence of a PLIF implies the existence of a universally measurable SPLIF?*

Question 4: *Does (γ) imply (β)? Equivalently, does the existence of a net of countable type of Borel functions which is consistent in probability imply the existence of a PLIF ?*

I guess the answer to the Question 4 is no. Finally, Lutz Weis [21] gave a mild justification of the intuition behind question 1 with the following

Theorem 4 *(Weis) The existence of a SPLIF is equivalent to the following finitely additive version of property $(\gamma_\perp)$:*

$(\alpha_\perp)$ For any two orthogonal finitely additive priors $\lambda_0 \perp \lambda_1$ on $(\Theta, \mathcal{T})$ the mixtures p_{λ_0} and p_{λ_1} are orthogonal as well.

****

Writing this review I experience once more the fascination by these questions which are simply put, relate easily to the most formal aspects of mathematics and at the same time help to clarify the way how to speak about

statistical concepts. In the mean time I think it would be interesting to understand more clearly how these different versions of a 'perfect' experiment could be approximated by finite-dimensional or even finite experiments. I believe in particular that a Shannon theoretic approach will be helpful in this endeavour.

Acknowledgement I am grateful to Dan Mauldin who read the first draft of this paper and made some useful suggestions.

References

[1] T. Bayes. An Essay Towards Solving a Problem in the Doctrine of Chances (1763). In *Facsimiles of two papers by Bayes*. Hafner, New York and London, 1963.

[2] D. Blackwell. There are no Borel SPLIFs. *Ann. of Prob.*, 8:1189–1190, 1980.

[3] D. Blackwell. A hypothesis-testing game without a value. *Festschrift for Erich L. Lehmann*, pages 79–82, 1983.

[4] D. Blackwell and R.V. Ramamoorthi. A Bayes but not classically sufficient statistic. *Ann. of Stat.*, 10:1025–1026, 1982.

[5] J.L. Doob. Application of the theory of martingales. *Coll. Int. du CNRS*, Paris:22–28, 1949.

[6] E.B. Dynkin. Sufficient statistics and extreme points. *Ann. of Prob.*, 6:705–730, 1978.

[7] A. Goullet de Rugy. Sur les mesures étrangères. *C.R. Acad. Sci. Paris*, 272:123–126, 1971.

[8] S. Graf and G. Mägerl. Families of pairwise orthogonal measures. In *9th Winter School in Abstract Analysis, Math. Inst. of the Cz. Acad. of Sciences, Praha*, 1981.

[9] G. Grimeisen. Ein Approximationssatz für Bairesche Funktionen. *Math. Ann.*, 146:189–194, 1962.

[10] N.P. Jewell. Mixtures of exponential distributions. *Ann. of Stat.*, 10:479–480, 1982.

[11] M. Katetov. On descriptive classes of functions. In G. Asser, J. Flachsmeyer, and W. Rinow, editors, *Theory of Sets and Topology*, pages 265–278. Deutscher Verlag der Wissenschaften, Berlin, 1972.

[12] K. Krickeberg. The Cox process. *Sympos. Math.*, 9:151–167, 1972.

[13] D. Maharam. Orthogonal measures: an example. *Ann. of Prob.*, 10:879–880, 1982.

[14] R.D. Mauldin, D. Preiss, and H.v. Weizsäcker. Orthogonal Transition Kernels. *Ann. of Prob.*, 11:970–988, 1983.

[15] R.D. Mauldin and H.v. Weizsäcker. Some orthogonality preserving kernels which are not completely orthogonal. *Ann. of Prob.*, 19:396–400, 1991.

[16] D. Ornstein and B. Weiss. How sampling reveals a process. *Ann. of Prob.*, 18:905–930, 1990.

[17] S. Shelah. Can you take Solovay's Inaccessible away? *Israel J. Math.*, 48,No.1:1–47, 1984.

[18] R. Solovay. A model of set theory in which every set of reals is Lebesgue measurable. *Ann. of Math.*, 92:1–56, 1970.

[19] M. Talagrand. Separation of orthogonal sets of measures (result of Mokobodzki). *9th Winter School in Abstract Analysis, Math. Inst. of the Cz. Acad. of Sciences, Praha*, 1981.

[20] J. Štepan. The probability limit identification function exists under the continuum hypothesis. *Ann. of Prob.*, 1:712–715, 1973.

[21] L.W. Weis. A Characterization of Orthogonal Transition Kernels. *Ann. of Prob.*, 12:1224–1227, 1984.

[22] H.v. Weizsäcker and G. Winkler. Non compact extremal integral representations: Some probabilistic aspects. *In: Functional Analysis: Surveys and recent results II, North Holland Publishers*, pages 115–148, 1980.

Heinrich v. Weizsäcker
Fachbereich Mathematik der Universität
Postfach
D 67663 Kaiserslautern
Germany
e-mail: weizsaecker@mathematik.uni-kl.de