Model-Free Data Condensation

Peter D. Finch

Phil. Trans. R. Soc. Lond. A 1991 **335**, 1-50
doi: 10.1098/rsta.1991.0035

Model-free data condensation

By Peter D. Finch

Department of Mathematics, Monash University, Clayton, Melbourne, Victoria, Australia 3168

Contents

Phil. Trans. R. Soc. Lond. A (1991) **335**, 1–50

Printed in Great Britain

1

Vol. 335. A (15 April 1991)

One aim of data analysis is its condensation, namely capturing its gist in an apposite way. This paper addresses the problem of constructing and assessing such condensations without reference to mechanisms which might have generated the data. The results obtained lead to non-probabilistic interpretations of some well-known inferential procedures of classical statistics and thereby shed new light on the structure of statistical inference and the theory of probability.

1. Introduction

Condensation of data is the suppression of its complex fine detail in favour of a simplified summarizing description. For our purposes an N-ary data-set is a finite ordered list $D = x_1, x_2, \ldots, x_N$ of not necessarily distinct readings; one for each of N cases, x_n being the reading for case n. Readings are elements of a possibly infinite set R which has a context-dependent structure, typically they are finite real vectors. The frequency function of the data-set D is the function F_D on R for which $F_D(x)$ is the number of times x in R occurs in the list defining D. The data support S_D is the set of distinct readings in D; it is also the support of F_D, namely the set of x in R at which $F_D(x)$ is not 0. For brevity we call F_D the spectrum of D; it sums to N, the arity of D. The normalization of F_D is the function f_D on R with $f_D(x) = N^{-1} F_D(x)$, it is called the density of D; it has the same support as F_D.

Let R_N be the set of N-ary data-sets, let R_* be the union of the R_N for $N \geqslant 1$ and let M be a non-empty set. A function $\delta: R_* \rightarrow M$ is called a condensing statistic. It condenses data D to $\delta(D)$ in M and this condensation can be used to describe D. An important condensing statistic is the function δ_σ whose value $\delta_\sigma(D)$ at D in R_* is its spectrum F_D. This condensing statistic leads to the familiar description of data by its frequency distribution when we are not concerned with which cases exhibit which readings. It suppresses the linkage between the readings and the cases from which they originate. We then specify data-sets only up to the equivalence determined by equality of distribution and do not distinguish between data-sets with the same spectrum. We call this the standard context because we use it as a benchmark when, more generally, we condense data to a possibly vector-valued function

$$\delta(D) = \Delta(F_D) \tag{1.1}$$

of its spectrum. Situations in which this type of condensation is used will be called macrostandard contexts; in them, we specify data-sets only up to equality of condensation and, for the purposes of the enquiry then at hand, we do not distinguish between data-sets with the same condensation.

When we condense data A to $\Delta(F_A)$ we record that it is one of the data-sets D with $\Delta(F_D) = \Delta(F_A)$. One who knows only that condensation cannot point to which of those data-sets is the actual data-set, because they all have the same condensation. Similarly one cannot, in general, recover F_A from the condensation $\Delta(F_A)$. Although F_A could be recovered by returning to A itself, to do so would negate the point of its suppression, namely the replacement of the intangible complexity of F_A by something simple and tangible. In these circumstances it is fruitful to talk about the macrostandard context as if it were a standard one with a known spectrum. To do so we use a surrogate spectrum $\tilde{F}_A$ in place of the suppressed actual spectrum F_A and talk about the macrostandard context as if it were a standard one with spectrum $\tilde{F}_A$. Expressions involving F_A are then replaced by the ones obtained from them by substituting $\tilde{F}_A$ for F_A. More familiar terms for such substitutions are estimation, approximation and their cognates; but these involve connotations of proximity to actuality which we wish to separate from surrogation *per se*, namely the idea of using a stand-in apart from the question of how well it plays that role, and our use of the neutral term 'surrogate' emphasizes this. By definition, a surrogate spectrum is a non-negative function on R with finite support, its normalization is the associated surrogate density.

In what follows we examine grounds for preferring certain types of surrogate spectra and develop a procedure which leads to a unique surrogate spectrum for use in a given macrostandard context. It turns out that this procedure is similar to ML-estimation and this fact leads to purely descriptive interpretations of some well-known inferential procedures which are commonly seen as taking their meaning from stochastic models for generating data. To avoid full generality at the outset, we first develop our argument in special projective contexts for which relatively few general concepts are required.

We consider only macrostandard contexts, namely data condensations of the form (1.1). This rules out situations in which the case-order in the data-set is contextually important, for instance when cases are epochs and the data-set is an individual time series. On the other hand we do have a macrostandard context when each reading is an individual time series and we are not concerned with which case is linked to which series, in other words when we have a finite ensemble of time series.

2. Macrolevels and their representation

Let δ be a condensing statistic. The equivalence relation ρ for which $D'\rho D''$ when $\delta(D') = \delta(D'')$ is said to be the macrolevel determined by δ. The ρ-equivalence class containing D is denoted by $\rho(D)$ and called the D-macrodatum. Many different condensing statistics determine the same macrolevel and each of them is said to represent it. A statistic which represents ρ has the form $\delta(D) = W\{\rho(D)\}$, where W is a one-to-one function on ρ-classes; we say that $\delta(D)$ is the name of $\rho(D)$ in the δ-representation. The macrolevel of the standard context is the symmetry relation σ on R_* for which $D'\sigma D''$ means that D' and D'' are permutations of each other. The macrolevels of the macrostandard contexts are the equivalences ρ on R_* which contain the symmetry σ. A macrostandard context with macrolevel ρ is called a ρ-context.

In practice there are often contextual constraints which rule out some data-sets. We suppose that these are constraints on data spectra that restrict discussion to the data-sets in a specified subset $\mathscr{C}$ of R_*, which is such that two data-sets with the same spectrum are either both in $\mathscr{C}$ or both not in $\mathscr{C}$. We treat the constraint set $\mathscr{C}$ as part of the data condensation by replacing δ of (1.1) with δ^* given by

$$\delta^*(D) = \{\delta(D), I(F_D)\},$$

where I is the indicator function of the set $\{F_D : D \in \mathscr{C}\}$. We call δ^* the contextual version of δ and the macrolevel it determines is called the contextual macrolevel. In what follows it is taken as understood that the condensing statistic is the contextual one when there are constraints.

A context defined in terms of a suggested condensing statistic, a proposed set of readings R and a given constraint set $\mathscr{C}$ sometimes turns out to be the standard context. This occurs when the function Δ of (1.1) is invertible on the constrained set of spectra so that $F_D = \Delta^{-1}\{\delta(D)\}$ is then recoverable from its condensation. Such a context is said to be identifiable.

For simplicity it is convenient to suppose that we know the support S_A of the suppressed spectrum F_A. There is no loss of generality in this supposition because we can partition R_* into disjoint cells each of which consists of all the data-sets with a corresponding particular common finite support and then consider each cell separately. When S_A is not in fact known it may be regarded as a generic support corresponding to a cell of R_*. We also suppose that the surrogate spectra $\tilde{F}_A$ which are under consideration as possible stand-ins for F_A all have the same support as F_A. In what follows, therefore, we consider an arbitrary macrostandard context with a known finite support S which is the common support of all the data-sets, spectra and surrogate spectra under consideration. The set of all data-sets with support S is denoted by $\mathscr{D}$, $\mathscr{F}$ is the corresponding set of spectra and $\mathscr{G} \supset \mathscr{F}$ is the set of surrogate spectra with support S. The size of S is denoted by V and $\mathscr{V}$ is the V-dimensional vector space of real functions on S equipped with the euclidean metric and orthogonality relation $\perp$ based on the inner product

$$(a, b) = \sum_S a(x)\, b(x). \tag{2.1}$$

A real function on R with support S can be regarded as an element of $\mathscr{V}$ by identifying it with its restriction to S. In this way all the spectra and surrogate spectra in $\mathscr{G}$, together with their densities, can be regarded as elements of $\mathscr{V}$ and, since they are everywhere positive on S, their logarithms are also in $\mathscr{V}$. Conversely a vector θ in $\mathscr{V}$ which is not zero anywhere on S can be regarded as a function on R with support S, by defining $\theta(x)$ to be 0 when x in R is not in S. We adopt both of these viewpoints and move from one to the other without further comment.

If Φ is a subspace of $\mathscr{V}$ and $\mathbb{P}$ is projection onto Φ, then the condensing statistic

$$\delta(D) = \mathbb{P}F_D \tag{2.2}$$

plays a central role in the following discussion. We call it a projection and the macrostandard context it determines is said to be a projective context. Its macrolevel is the equivalence π given by

$$D' \pi D'' \Leftrightarrow F_{D'} - F_{D''} \perp \Phi. \tag{2.3}$$

This macrolevel can also be represented in terms of linearly independent vectors $\phi_0, \phi_1, \ldots, \phi_M$ spanning $\mathit{\Phi}$ by means of the vector-valued condensing statistic

$$\phi(D) = (\phi_0(D), \phi_1(D), \ldots, \phi_M(D)), \tag{2.4}$$

where

$$\phi_m(D) = (\phi_m, F_D) = \sum_{n=1}^{N} \phi_m(x_n), \tag{2.5}$$

with $D = x_1 x_2 \ldots x_N$.

When $\mathit{\Phi}$ contains the constant vector 1, namely the element of $\mathscr{V}$ which is identically 1 on S, it follows from (2.3) that $(1, F_{D'}) = (1, F_{D''})$; in other words data-sets with the same condensation then have the same arity. Such a projection is said to be arity restricted. The representation (2.4) is said to be regular when (i) $\phi_0(x) \equiv 1$ on S, so that $\phi_0(D) = N$ is the arity of D and the projection is arity restricted, and when (ii) $\phi_m \perp 1$ for each $m = 1, 2, \ldots, M$.

3. Consistent surrogation

In a ρ-context with the representing statistic δ of (1.1) the suppressed spectrum F_A of the actual data-set A is condensed to $\Delta(F_A)$. If $\tilde{F}_A$ is a surrogate spectrum used in place of F_A, then the substitutional surrogate for the condensation is $\Delta(\tilde{F}_A)$. The surrogate $\tilde{F}_A$ is said to be consistent at A when

$$\Delta(\tilde{F}_A) = \Delta(F_A). \tag{3.1}$$

When the context is identifiable Δ is invertible and (3.1) gives $\tilde{F}_A = F_A$. Consistency does not depend on the representing statistic because each of them is a one-to-one function on ρ-classes. Equation (3.1) is called the A-designation equation of the ρ-context. In what follows we require that the surrogates used are consistent, namely satisfy (3.1) for the data-set A under consideration, partly because it is intuitively plausible to do so and partly because this has interesting consequences. We do not argue that there are normative grounds for adopting (3.1).

Although it seems sensible to use consistent surrogates there is a technical difficulty with the designation equation (3.1). For the condensing statistic δ is given by (1.1) in terms of a function Δ which is defined on data spectra, but for the designation equation to have meaning Δ must also be defined on surrogate spectra. In other words we have to extend Δ from $\mathscr{F}$ to $\mathscr{G}$. There is no difficulty when the result of substituting $\tilde{F}_A$ for F_A in the expression $\Delta(F_A)$ has an unambiguous mathematical meaning as, for example, in a projective context. For if δ is the projection (2.2), then we can extend Δ from $\mathscr{F}$ to $\mathscr{G}$ by defining it on $\mathscr{G}$ through the expression

$$\Delta(G) = \mathbb{P}G, \forall\, G \in \mathscr{G}.$$

Adopting this extension of Δ, the designation equation (3.1) takes the unambiguous form

$$\mathbb{P}\tilde{F}_A = \mathbb{P}F_A. \tag{3.2}$$

In the representation of the statistic ϕ of (2.4) this equation takes the form

$$(\phi_m, \tilde{F}_A) = (\phi_m, F_A), \quad 0 \leqslant m \leqslant M. \tag{3.3}$$

But in a context defined in terms of an arbitrary macrolevel the corresponding extension of Δ is not so obvious. We return to this difficulty later. For the time being we discuss only projections. The general case is addressed in §5.4.

4. Surrogation in projective contexts

In this section we consider arity restricted projective contexts and certain exponential families of surrogate spectra associated with them. We show that the designation equation is the ML-estimation equation in that family and consider a number of issues raised by this fact. We present a descriptive version of Bayes's postulate, a descriptive interpretation of the concept of probability in binary sequences and illustrate explanatory surrogation by means of a generalized linear model. Finally we discuss entropy, likelihood and their maximizations.

4.1. *Exponential families*

For each θ in $\mathscr{V}$, exp θ is also in $\mathscr{V}$ and, since it is everywhere positive on S, it is an element of $\mathscr{G}$. Conversely each surrogate spectrum G in $\mathscr{G}$ arises in this way because $\theta = \ln G$ is in $\mathscr{V}$. Thus any non-empty subset Γ of $\mathscr{G}$, including $\mathscr{G}$ itself, can be presented as the exponential family

$$\mathscr{E}(\boldsymbol{\Theta}) = \{\exp \gamma : \gamma \in \boldsymbol{\Theta}\}, \quad \boldsymbol{\Theta} = \ln \Gamma. \tag{4.1}$$

The exponential family is said to be flat when $\boldsymbol{\Theta}$ is a subspace of $\mathscr{V}$ and is said to be curved otherwise. The densities corresponding to the surrogate spectra to $\mathscr{E}(\boldsymbol{\Theta})$ are the functions

$$g(x \mid \theta) = \mathrm{e}^{\theta(x)}/(1, \mathrm{e}^{\theta}) = \exp\{\theta(x) - \ln Z(\theta)\}, \tag{4.2}$$

where

$$Z(\theta) = \sum_S \mathrm{e}^{\theta(x)} = (1, \mathrm{e}^{\theta}). \tag{4.3}$$

In (4.1) vectors θ', θ'' which differ by a constant vector of $\mathscr{V}$ determine the same density in (4.2). The exponential family of densities $g(x \mid \theta)$ with θ in $\boldsymbol{\Theta}$ is denoted by $e(\boldsymbol{\Theta})$. The following result is worth noting.

Theorem 4.1. *Any set* $\{g_m : m \in M\}$ *of densities with common support* S *form the exponential family* $e(\boldsymbol{\Theta}_M)$, *where* $\boldsymbol{\Theta}_M \perp 1$ *is the set of vectors*

$$\theta_m(x) = -V^{-1}(1, \ln g_m) + \ln g_m(x), \quad m \in M. \tag{4.4}$$

Proof. With $\theta_m(x)$ given by (4.4), $g(\cdot \mid \theta_m)$ of (4.2) is $g_m(x)$ and $(1, \theta_m) = 0$.

4.2. *Designation and maximum likelihood*

Let δ of (2.2) be an arity restricted projection and let ϕ of (2.4) be a regular representation of it. The first of the designation equations (3.3), corresponding to $m = 0$, is $(1, \tilde{F}_A) = N$ where $N = \phi_0(D)$, the arity of A, is part of the data condensation $\phi(D)$. Dividing each of the remaining M equations of (3.3) by the first of them, we obtain the designation equation in the ϕ-representation in terms of densities, namely

$$(\phi_m, \tilde{f}_A) = (\phi_m, f_A), \; 1 \leqslant m \leqslant M, \tag{4.5}$$

or equivalently

$$\mathbb{P}_* \tilde{f}_A = \mathbb{P}_* f_A, \tag{4.6}$$

where $\mathbb{P}_*$ is projection onto the subspace $\boldsymbol{\Phi}_* = \boldsymbol{\Phi} \cap 1^{\perp}$.

Consider the exponential family of densities $e(\boldsymbol{\Phi}_*)$, namely those of the form

$$f^{\phi}(x) = \exp\{\phi(x) - \ln Z(\phi)\}, \quad \phi = \sum_{m=1}^{M} q_m \phi_m \in \boldsymbol{\Phi}_*. \tag{4.7}$$

It is well-known that if we regard f^{ϕ} as the density of a population from which the data A was obtained by ordered random sampling with replacement, then the partial derivatives of the log-likelihood function l_A^{ϕ} are

$$\partial l_A^{\phi}/\partial q_m = N\{(\phi_m, f_A) - (\phi_m, f^{\phi})\}, \tag{4.8}$$

where N, the arity of A, is the sample size. Comparison of (4.5) and (4.8) shows that if we restrict ourselves to the use of the surrogate spectra in $e(\varPhi_*)$, then the solution to the designation equations (4.5) is the ML-estimate within that exponential family.

Thus if we adopt consistency and use only surrogate densities in $e(\varPhi_*)$, then there is a uniquely determined density designated for use in place of f_A, namely the density $\hat{f}_A$ which is the ML-estimate of the population density under the assumptions that (i) the population density is in $e(\varPhi_*)$ and (ii) the data A is an ordered random sample from that population. In our framework, however, the designated density $\hat{f}_A$ is a surrogate for the 'sample' density suppressed by the data condensation; it is not considered as an estimate of an underlying population density. Moreover maximization of likelihood is not advanced as the reason for using $\hat{f}_A$, it arises as a consequence of consistency in the special exponential family $\mathscr{E}(\varPhi)$. Reasons for choosing our surrogate spectra from that family are examined first in §5.2 and then more deeply in §9.

In classical statistics there are pathological situations in which there is no ML-estimate $\hat{f}_A$ in the family $e(\varPhi_*)$. But in the descriptive framework presented here there is always a unique solution in $e(\varPhi_*)$ to the designation equation (4.6) and the analogue of the classical pathology does not arise because all the densities under consideration have the same support. For completeness we discuss this fact in the next section.

Equation (4.7) presents $e(\varPhi_*)$ in the form of an exponential family of probability densities admitting the $\phi_m(A) = (\phi_m, F_A)$ as sufficient statistics. A more familiar form of that family arises when the $\phi_m(x) = \theta_m\{T(x)\}$ for some functions θ_m and T. If the surrogate density f^{ϕ} of (4.7) is then used in place of f_A, and if t is in the range of the function T, then the corresponding surrogate for the relative frequency of readings x in the data set A with $T(x) = t$ is the sum of the $f^{\phi}(x)$ over those readings, namely

$$f_T^{\phi}(t) = C(\phi)\, h(t) \exp\left\{\sum_{m=1}^{M} q_m\, \theta_m(t)\right\},$$

where $h(t)$ is the proportion of x in S with $T(x) = t$ and $C(\phi)$ is a normalizing constant. This is a standard form for an exponential family of probability densities. Thus using only the surrogate spectra in $\mathscr{E}(\varPhi)$ can be viewed as modelling the suppressed data density by the exponential family of densities for which its projective condensation is sufficient in the sense of classical statistics. But while this relates surrogation in $\mathscr{E}(\varPhi)$ to more familiar practices it is not a justification for using only surrogates from that family.

Each surrogate spectrum G in $\mathscr{E}(\varPhi)$ is such that

$$\forall x \in S : \ln G(x) = \phi(x), \quad \phi\, \mathscr{E}\, \varPhi$$

and can be viewed as a log-linear representation of the suppressed data spectrum in which ϕ in $\varPhi$ plays the role of a vector-valued explanatory variable.

4.3. *The existence theorem*

The existence of a unique solution in $e(\Phi_*)$ to the designation equation (4.6) follows from general results in Rockafellar (1970) and Barndorff-Nielson (1978), here we prove this by elementary arguments. We start with two lemmas.

Lemma 4.1. *The maximum and the minimum of* (ψ, f_A) *for* ψ *in* $\mathscr{V}$ *with* $\|\psi\| = 1$ *and* $\psi \perp 1$ *are* $\pm d$ *where*

$$d^2 = \sum_S \left(f_A(x) - \frac{1}{V}\right)^2, \quad d > 0.$$

If f_A *is not constant, then the maximum and minimum are obtained at* $\pm\psi_*(x)$ *where*

$$\psi_*(x) = d^{-1}\{f_A(x) - V^{-1}\}.$$

Proof. If f_A is constant, then $d = 0$ and $(\psi, f_A) \equiv 0$ for $\psi \perp 1$. When f_A is not constant, $d \neq 0$ and $\|\psi\| = 1$, $\psi \perp 1$ give

$$\tfrac{1}{2}d\,\|\psi+\psi_*\|^2 - d = (\psi, f_A) = d - \tfrac{1}{2}d\,\|\psi - \psi_*\|^2,$$

and the lemma follows.

Lemma 4.2. *For each* $\psi \neq 0$ *in* Φ_*, *there is exactly one real number* $a = a(\psi)$ *such that*

$$(\psi, \mathrm{e}^{a\psi})/(1, \mathrm{e}^{a\psi}) = (\psi, f_A). \tag{4.9}$$

Moreover, if $\|\psi\| = 1$, *then there exist numbers* H *and* K *which do not depend on* ψ *such that*

$$H \leqslant a(\psi) \leqslant K. \tag{4.10}$$

Proof. Let l and L be the minimum and maximum of ψ on S. Since $\psi \perp 1$, and is not 0, it is not constant and so $l < L$. Since f_A is strictly positive on S,

$$l < (\psi, f_A) < L. \tag{4.11}$$

For each real c write

$$W(c) = (\psi, g^{c\psi}) \tag{4.12}$$

for the mean of ψ with respect to the density $g^{c\psi} = \mathrm{e}^{c\psi}/(1, \mathrm{e}^{c\psi})$. Let λ and Λ be the subsets of S where $\psi = l$ and $\psi = L$ respectively, with sizes $|\lambda|$, $|\Lambda|$, and let λ', Λ' be their set complements in S. Then

$$W(c) - l = \sum_{\lambda'}(\psi - l)\,\mathrm{e}^{c(\psi - l)}/\{|\lambda| + \sum_{\lambda'}\mathrm{e}^{c(\psi - l)}\},$$

$$L - W(c) = \sum_{\Lambda'}(L-\psi)\,\mathrm{e}^{c(L-\psi)}/\{|\Lambda| + \sum_{\Lambda'}\mathrm{e}^{c(L-\psi)}\}.$$

Thus
$$l = W(-\infty) < W(c) < W(+\infty) = L. \tag{4.13}$$

But $W'(c)$ is the variance of ψ with respect to the density $g^{c\psi}$ and it is positive because ψ is not constant. Thus $W(c)$ is strictly increasing in c and, from (4.13), it takes on all values between l and L. It follows from (4.11) that there is just one value of c, a say, satisfying (4.9). Finally, Lemma (4.1) shows that if $\|\psi\| = 1$, then $|W(a)| = |(\psi, f_A)| \leqslant d$ and so we have (4.10) with $H = W^{-1}(-d)$ and $K = W^{-1}(d)$.

Remark. The key fact in proving the existence of a finite solution a to (4.9) is the strict positivity of f_A on S. This rules out the possibility of equality at either of the bounds in (4.13). But if, for example, f_A was only required to be non-negative on S

and $\psi = L$ where $F_A \neq 0$, then $(\psi, f_A) = L$ and there would be no finite a satisfying (4.9). That equation would only be satisfied 'at infinity', in the sense that $g^* = \lim g^{c\psi}$ is $|A|^{-1} I_A$, where I_A is the indicator function of A and $(\psi, g^*) = L = (\psi, f_A)$. This sort of pathology cannot arise here because the data densities under consideration are all strictly positive on S.

When $M = \dim \Phi_* = 1$, Lemma (4.2) establishes the existence of a unique solution in $e(\Phi_*)$ to the designation equation (4.6). When $M > 1$ put

$$Q(\psi) = \ln\,(1, \mathrm{e}^{\psi}) - (\psi, f_A), \quad \psi \in \Phi_*. \tag{4.14}$$

If $\psi = a_1 \phi_1 + a_2 \phi_2 + \ldots + a_M \phi_M$, then $\partial Q(\psi)/\partial a_m$ is $(\phi_m, g^{\psi}) - (\phi_m, f_A)$ and so the solutions to equations (4.5) in $e(\Phi_*)$ correspond to stationary points of Q. Standard arguments show that the hessian of Q is everywhere positive definite. Thus Q is strictly convex and its stationary points are minima. The existence of a unique global minimum, and hence of a unique density in $e(\Phi_*)$ satisfying the designation equations (4.5), follows from the fact that the local minima all lie in a bounded convex region; this is an immediate consequence of Lemma (4.2). For

$$\frac{\mathrm{d}}{\mathrm{d}c} Q(c\psi) = W(c) - (\psi, f_A), \quad \frac{\mathrm{d}^2}{\mathrm{d}c^2} Q(c\psi) = W'(c) > 0,$$

where $W(c)$ is given by (4.12). Thus the solution $a = a(\psi)$ to (4.9) determines the unique vector $a(\psi)\,\psi$ at which $Q(c\psi)$ attains its minimum in the direction of ψ. Let $\Omega = \{a(\psi)\,\psi : \psi \in \Phi_*\ \&\ \|\psi\| = 1\}$ be the set of all these vectors. By (4.10), the $|a(\psi)|$ have a common finite upper bound and so all the local minima of Q lie in a bounded convex region of Φ_*, for instance a large enough ball centred at 0.

4.4. *Bayes's postulate*

The simplest arity restricted projection is the one determined by the subspace $\Phi = [1]$, the one-dimensional subspace of constant vectors. The data is then condensed to its arity, Φ_* contains only the 0 vector and there is only one density in $e(\Phi_*)$, namely the uniform one $\hat{f}_A(x) \equiv 1/V$ on S, this being the designated density because equations (4.5) are vacuous in this case. Condensation to arity can be seen as suppressing all the information in the data except for the number of case-readings in it. Interpreting the suppression of information as its absence, the procedure suggested in §4.2 leads to the following descriptive version of Bayes's postulate:

> If a finite set of case-readings is condensed to its arity, then their surrogate density is uniform on the data support.

This version of the postulate does not assert that the actual data density is uniform, only that the suggested surrogation procedure leads to the uniform surrogate for that density when the data is condensed in a very special way. Moreover it is implicit in the statement of it that there are no contextual constraints, because if there were, then the contextual condensing statistic would give more than just the arity of the data-set.

4.5. *Binary sequences*

Suppose that each reading in the data-set A is the record of a binomial success/failure experiment consisting of n trials. The data support S is then a finite set of binary sequences of length n and its size $V \leqslant \min\,(N, 2^n)$ where N, the arity of

A, is the number of experiments. Suppose also that we condense the data set A to (i) its arity N and (ii) the total number of successes in all the Nn trials. This condensation is an arity restricted projection which has the representation (2.4) with $M = 1$, $\phi_0 \equiv 1$ on S and $\phi_1(x)$ the number of successes in the reading x. From (2.5) the condensation of A in this representation consists of $\phi_0(A) = N$, the arity of A, and $\phi_1(A) = T$, the total number of successes in all the Nn trials. The overall success rate is T/Nn.

The densities $g(x)$ in $e(\Phi_*)$ have the form $K(q)\exp\{q\phi_1(x)\}$, where q is a real number and $K(q)$ is a normalizing constant. Writing p for $\exp(q)/\{1+\exp(q)\}$ and $n(x)$ for $\phi_1(x)$, we have

$$g(x) = Cp^{n(x)}(1-p)^{n-n(x)}, \quad x \in S, \tag{4.15}$$

where $C = C(p)$ is a normalizing constant. This is the generic form of the densities in $e(\Phi_*)$. The designation equation (4.5) is

$$C\sum_S n(x)\,p^{n(x)}(1-p)^{n-n(x)} = T/N. \tag{4.16}$$

Writing $S(t)$ for the number of x in S with exactly t successes, equation (4.16) takes the form

$$\sum_{t=0}^{n} tS(t)\,p^t(1-p)^{n-t} \Big/ \sum_{t=0}^{n} S(t)\,p^t(1-p)^{n-t} = T/N. \tag{4.17}$$

In this equation the numbers $S(t)$ are supposed known because it is assumed that the data support is a known set. The unique $p = \hat{p}$ satisfying the designation equation (4.17) can be determined by standard iterative procedures. Thus the designated surrogate density has the form

$$\hat{f}_A(x) = C(\hat{p})\,\hat{p}^{n(x)}(1-\hat{p})^{n-n(x)}, \quad x \in S. \tag{4.18}$$

When $n = 1$ the context is identifiable. If $n \geqslant 2$ and S consists of all the 2^n binary sequences of length n, in which case $N \geqslant V = 2^n$, then $S(t)$ is the binomial coefficient $\binom{n}{t}$. In this case $C(p) \equiv 1$ and the solution to (4.17) is $\hat{p} = T/Nn$, the overall success rate, and

$$\hat{f}_A(x) = \hat{p}^{n(x)}(1-\hat{p})^{n-n(x)}, \quad \hat{p} = T/Nn, \tag{4.19}$$

for each x in S. In particular the corresponding surrogate for the relative frequency in the data set A of component experiments with exactly t successes is

$$\hat{f}_A\{n^{-1}(t)\} = \binom{n}{t}\hat{p}^t(1-\hat{p})^{n-t}, \quad 0 \leqslant t \leqslant n, \tag{4.20}$$

namely the binomial probability for t successes in n independent trials with common success probability $\hat{p} = T/Nn$, the overall success rate.

The condensation of a finite number of equally sized finite binary sequences to their arity and total successes is not fruitful when the surrogate $\hat{f}_A$ of (4.18) does not mimic the suppressed actual density f_A. It is fruitful for randomly generated independent trials because, in a very large series of such experiments, the $\hat{f}_A(x)$ of (4.19) and the $\hat{f}_A\{n^{-1}(t)\}$ of (4.20) are likely to be close to their sample counterparts, namely the suppressed data densities $f_A(x)$ and $f_A\{n^{-1}t)\}$. On the other hand, the large N requirement together with (4.19) is suggestive of the way the concept of probability is related to long-run relative frequency in certain types of binomial sequences. Looking from data analysis to probability theory, rather than in the reverse direction, probability in statistical analysis can be seen as a convenient paraphrase of the fact that, through condensation to arity and total successes, the overall success rate provides fruitful descriptions of data generated by simple games of chance and the adoption of a paradigm which sees most, if not all, data as generated, at least conceptually, by similar mechanisms.

4.6. *Explanatory surrogation and generalized linear models*

Data analysis is often directed to the examination of relationships between variables. For instance, suppose that each reading x is an ordered pair (ξ, η) of possibly vector-valued quantities and that g is a surrogate for f_A based on the condensation of the data-set A at the macrolevel ρ. Given ξ, the surrogate for the conditional density of η is

$$g(\eta \,|\, \xi) = g(\xi, \eta)/g(\xi, \cdot), \tag{4.21}$$

where $g(\xi, \cdot)$ is the marginal surrogate density for ξ. Equation (4.21) exhibits the surrogate dependence of the variable η on the variable ξ regarded as an explanatory case profile. We refer to its use in that way as explanatory surrogation.

When the condensing macrolevel ρ is the projection π of (2.3), the conditional densities derived from the surrogate spectra in $\mathscr{E}(\boldsymbol{\Phi})$ are the

$$g(\eta \,|\, \xi) = E(\xi, \eta)/E(\xi, \cdot), \tag{4.22}$$

where
$$E(\xi, \eta) = \exp \phi(\xi, \eta), \quad \phi \in \boldsymbol{\Phi}, \tag{4.23}$$

and $E(\xi, \cdot)$ is the sum of the $E(\xi, \eta)$ over those η for which (ξ, η) is in the data support S. If $\phi_k, 1 \leqslant k \leqslant K$, are linearly independent vectors spanning $\boldsymbol{\Phi}$, then

$$\phi(\xi, \eta) = \sum_{k=1}^{K} c_k \phi_k(\xi, \eta), \tag{4.24}$$

where the c_k are uniquely determined by ϕ.

To illustrate explanatory surrogation and relate it to generalized linear models, we condense the data by a projection which can be split into three parts. The first condenses the bivariate spectrum $F_A(\xi, \eta)$ in a special way, the second condenses the marginal spectrum $F_A(\cdot, \eta)$ and the third condenses the marginal spectrum $F_A(\xi, \cdot)$. To formulate this data condensation let P be the set of distinct profiles ξ in the data, let Q be the set of distinct values taken by the explained variable η in the data-set and let the data-set itself be $A = x_1, x_2, \ldots, x_N$ with $x_n = (\xi_n, \eta_n)$.

Let $\alpha_u, 1 \leqslant u \leqslant U$ be linearly independent functions on P and let w be a non-constant weighting function on Q. The first part of the data condensation condenses the bivariate spectrum $F_A(\xi, \eta)$ to the U statistics

$$W_u = \sum_{n=1}^{N} \alpha_u(\xi_n)\, w(\eta_n), \quad 1 \leqslant u \leqslant U. \tag{4.25}$$

Let $\beta_i, 1 \leqslant i \leqslant I$, be I linearly independent functions on Q. The second part of the data condensation condenses the marginal spectrum $F_A(\cdot, \eta)$ to the I statistics

$$Y_i = \sum_{n=1}^{N} \beta_i(\eta_n), \quad 1 \leqslant i \leqslant I. \tag{4.26}$$

Finally let $\psi_j, 1 \leqslant j \leqslant J$, be J linearly independent functions on P. The third part of the data condensation condenses the marginal spectrum $F_A(\xi, \cdot)$ to the J statistics

$$X_j = \sum_{n=1}^{N} \psi_j(\xi_n), \quad 1 \leqslant j \leqslant J. \tag{4.27}$$

To formulate this data condensation as a projection, introduce the $I+J+U$ vectors of V given by $\phi_u(\xi, \eta) = \alpha_u(\xi)\, w(\eta)$, $1 \leqslant u \leqslant U$, $\phi_{U+i}(\xi, \eta) = \beta_i(\eta), 1 \leqslant i \leqslant I$, and $\phi_{U+I+j}(\xi, \eta) = \psi_j(\xi), 1 \leqslant j \leqslant J$. Let Φ be the subspace of V which is spanned by these vectors. Since

$$\left.\begin{aligned} (\phi_u, F_A) &= \sum_S \alpha_u(\xi)\, w(\eta)\, F_A(\xi, \eta) = W_u, \\ (\phi_{U+i}, F_A) &= \sum_Q \beta_i(\eta)\, F_A(\cdot\,, \eta) = Y_i, \\ (\phi_{U+I+j}, F_A) &= \sum_p \psi_j(\xi)\, F_A(\xi, \cdot) = X_j, \end{aligned}\right\} \tag{4.28}$$

it follows from (2.5) that condensing the data to the statistics $W_u, 1 \leqslant u \leqslant U, Y_i$, $1 \leqslant i \leqslant I$, and $X_j, 1 \leqslant j \leqslant J$, corresponds to data condensation by the projection of F_A onto Φ.

Taking $K = U+I+J$ in (4.24), equations (4.23) and (4.24) give

$$E(\xi, \eta) = \alpha(\xi)\, w(\eta) + \beta(\eta) + \beta_*(\xi), \tag{4.29}$$

where

$$\left.\begin{aligned} \alpha(\xi) &= \sum_{u=1}^{U} c_u\, \alpha_u(\xi), \\ \beta(\eta) &= \sum_{i=1}^{I} c_{U+i}\, \beta_i(\eta), \\ \beta_*(\xi) &= \sum_{j=1}^{J} c_{U+I+j}\, \psi_j(\xi). \end{aligned}\right\} \tag{4.30}$$

Thus the generic conditional density given by (4.22) is

$$g(\eta \mid \xi) = \frac{\exp\{\alpha(\xi)\, w(\eta) + \beta(\eta)\}}{\sum_\zeta \exp\{\alpha(\xi)\, w(\zeta) + \beta(\zeta)\}}, \tag{4.31}$$

where the summation in the denominator extends over those ζ for which (ξ, ζ) is in the data support S. If we use a consistent surrogate spectrum from $\mathscr{E}(\Phi)$, then the constants c_k, are determined by the designation equations (3.3), namely by equations (4.28) with F_A on the left replaced by $\tilde{F}_A$. The values so obtained are the usual ML-estimates of those constants.

To relate the form of (4.31) to generalized linear models consider the special case $w(\eta) \equiv \eta$ and note that the conditional surrogate mean of η, namely

$$\mu_{\eta|\xi} = \sum_Q \eta g(\eta \mid \xi) = M\{\alpha(\xi)\}, \tag{4.32}$$

is a function M of $\alpha(\xi)$. Its derivative $M'\{\alpha(\xi)\}$ is the variance of η with respect to the conditional density $g(\eta \mid \xi)$ and is therefore positive for a ξ associated with more than one value of η. For such a ξ, $M\{\alpha(\xi)\}$ is a strictly increasing function of $\alpha(\xi)$ and so it has an inverse $M^{-1} = H$ say. Given the profile ξ,

$$\alpha(\xi) = H(\mu_{\eta|\xi}) \tag{4.33}$$

is a function of the conditional surrogate mean of η. This equation presents H as the link function of a generalized linear model in canonical form which is based on the exponential family

$$g(\eta \mid \xi) = \exp\{\alpha(\xi)\,\eta + \beta(\eta) + \gamma(\xi)\}, \tag{4.34}$$

where $\exp\gamma(\xi)$ is a normalizing factor. If we retain the general weighting w, then we obtain a generalized linear model in non-canonical form, $g(\eta \mid \xi)$ has the form $\exp\{\alpha(\xi)\, w(\eta) + \beta(\eta) + \gamma(\xi)\}$ and $\alpha(\xi) = H(\mu_{w|\xi})$ is a function of the conditional mean of $w(\eta)$.

By way of illustration, suppose that η is a binary 0, 1 variable. Equation (4.32) is

$$M\{\alpha(\xi)\} = g(1 \mid \xi) = \frac{\exp\{\alpha(\xi) + \beta(1)\}}{\exp\{\beta(0)\} + \exp\{\alpha(\xi) + \beta(1)\}}.$$

The inverse function M^{-1} is a shifted version of the logit and (4.33) is the logistic regression

$$\operatorname{logit} g(1 \mid \xi) = \beta(1) - \beta(0) + \alpha(\xi) \tag{4.35}$$

expressing $g(1 \mid \xi)$ in terms of a constant and the explanatory variables $\alpha_u(\xi)$, $1 \leqslant u \leqslant U$.

The classical analogue of the explanatory surrogation (4.34) would involve independent random variables $Y_1, Y_2, \ldots, Y_T$ having exponential densities of the same form but with different parameters. The density for Y_t would be

$$f(y;\theta_t) = \exp\{ya(\theta_t) + b(y_t) + c(\theta_t)\}, \tag{4.36}$$

where θ_t is the associated value of the parameter. With each Y_t there would be associated a vector

$$\boldsymbol{x}_t = (x_{1t}, x_{2t}, \ldots, x_{Ut}) \tag{4.37}$$

of explanatory variables and the link function L would relate the mean μ_t of Y_t to them by an equation of the form

$$L(\mu_t) = \sum_{u=1}^{U} \gamma_u\, x_{ut}, \tag{4.38}$$

where the function L is monotone and hence invertible, and the γ_u are estimated by, for example, maximum likelihood. There are obvious analogies between (4.34) and (4.36), and between the roles played by the $\boldsymbol{x}_t$ and the $\alpha(\xi)$, with x_{tu} being the analogue of $\alpha_u(\xi)$. Nevertheless there are two important differences. Firstly $g(\eta \mid \xi)$ of (4.34) is a conditional density whereas $f(y;\theta_t)$ of (4.36) is an unconditional density. Secondly, in the classical framework a number of different link functions might be considered in conjunction with the exponential families (4.36).

An argument like that leading from (4.31) to (4.32) and (4.33) shows that there is a function M, with inverse $M^{-1} = H$, such that

$$\mu_t = M\{a(\theta_t)\}, \quad a(\theta_t) = H(\mu_t).$$

Thus (4.38) gives

$$a(\theta_t) = HL^{-1}\left\{\sum_{u=1}^{U} \gamma_u\, x_{ut}\right\}. \tag{4.39}$$

The natural link function $L = H$ arises in the explanatory surrogation considered here because we are using only the surrogate spectra in $\mathscr{E}(\boldsymbol{\Phi})$, where $\boldsymbol{\Phi}$ is the subspace determining the projective data condensation then under study. We return to this point in §9.4.

4.7. *Entropy and likelihood*

The entropy of a density g in $\mathscr{G}$ is

$$\operatorname{Ent}(g) = -(\ln g, g). \tag{4.40}$$

It is well known that if f and g are any two densities in $\mathscr{G}$, then

$$\operatorname{Ent}(g) \leqslant -(\ln f, g), \tag{4.41}$$

with equality only when $f = g$. This follows from the fact that, for positive real c, $\ln c \geqslant 1 - c^{-1}$; by taking $c = g(x)/f(x)$ and adding over the x in S.

Let $\boldsymbol{\Phi}$ be a subspace of $\mathscr{V}$ which contains the constant vector 1 and, as in §4.2, let $\boldsymbol{\Phi}_* = \boldsymbol{\Phi}_{\cap} 1^{\perp}$. We say that a density g in $\mathscr{G}$ is $\boldsymbol{\Phi}_*$-consistent at the data-set A when it has the same projection onto $\boldsymbol{\Phi}_*$ as f_A, namely when

$$(\psi, g) = (\psi, f_A), \forall \psi \in \boldsymbol{\Phi}_*. \tag{4.42}$$

By the existence theorem of §4.3, there is exactly one density in $e(\boldsymbol{\Phi}_*)$ which is $\boldsymbol{\Phi}_*$-consistent at A; let it be

$$\hat{f}_A = f^{\phi} = e^{\phi}/(1, e^{\phi}), \tag{4.43}$$

where ϕ is the unique vector in $\boldsymbol{\Phi}_*$ which determines $\hat{f}_A$. It follows from (4.42) that $(\phi, f^{\phi}) = (\phi, f_A)$ and hence that

$$\operatorname{Ent}(\hat{f}_A) = -(\ln \hat{f}_A, f_A). \tag{4.44}$$

From this equation, together with (4.41) and (4.42), we see that if g is any density of $\mathscr{G}$ which is $\boldsymbol{\Phi}_*$-consistent at A, then

$$\operatorname{Ent}(g) \leqslant -(\ln \hat{f}_A, g) = -(\ln \hat{f}_A, f_A) = \operatorname{Ent}(\hat{f}_A) \tag{4.45}$$

with equality only when $g = \hat{f}_A$. In other words we have the following result.

Theorem 4.2. *In an arity restricted projective context with suppressed data-set A and condensation by projection onto $\boldsymbol{\Phi}$, the uniquely determined density in $e(\boldsymbol{\Phi}_*)$ which is $\boldsymbol{\Phi}_*$-consistent at A maximizes entropy over the class of all the densities in $\mathscr{G}$ which are $\boldsymbol{\Phi}_*$-consistent at A.*

The likelihood function of the data-set $A = x_1, x_2, \ldots, x_N$ is defined to be the function $\operatorname{lik}(\cdot \mid A)$ on $\mathscr{G}$ given by the expression

$$\operatorname{lik}(g \mid A) = \prod_{n=1}^{N} g(x_n) = \exp N(\ln g, f_A). \tag{4.46}$$

This is the likelihood function of classical statistics when A is regarded as an ordered random sample of size N drawn with replacement from a population with density g. From the earlier results of this section we obtain

$$\operatorname{lik}(g \mid A) \leqslant \exp\{-N \operatorname{Ent}(f_A)\} = \operatorname{lik}(f_A \mid A), \tag{4.47}$$

with equality only when $g = f_A$, and

$$\operatorname{lik}(\hat{f}_A \mid A) = \exp\{-N \operatorname{Ent}(\hat{f}_A)\}. \tag{4.48}$$

It follows from the results of §§4.2 and 4.3 that we have the following theorem.

Theorem 4.3. *In an arity restricted projective context with suppressed data-set A and condensation by projection onto Φ, the uniquely determined density in $e(\Phi_*)$ which is Φ_*-consistent at A maximizes likelihood over all the densities in $e(\Phi_*)$.*

This follows from (4.14), namely $Q(\psi) = -N^{-1} \ln \text{lik } (f^\psi | A)$. It can also be seen by writing $l_A(\cdot)$ for the log-likelihood $\ln \text{lik } (\cdot | A)$ and noting that, for any θ, ψ in Φ_*,

$$N^{-1}\{l_A(f^\psi)-l_A(f^\theta)\} = (\psi-\theta, f_A)+\ln\{(1, e^\theta)/(1, e^\psi)\}.$$

But (4.41) with $g = f^\psi$ and $f = f^\theta$ shows that the last term on the right is bounded below by $(\theta-\psi, f^\psi)$ and so

$$N^{-1}\{l_A(f^\psi)-l_A(f^\theta)\} \geqslant (\theta-\psi, f^\psi-f_A).$$

If $f^\psi = \hat{f}_A$, then the lower bound on the right is 0 by (4.42), because $\hat{f}_A$ is consistent at A and $\theta-\psi$ is in Φ_*. Thus $l_A(\hat{f}_A) \geqslant l_A(f^\theta)$ for all θ in Φ_* with equality only when $f^\theta = \hat{f}_A$.

Thus consistency in surrogation leads one to designate a surrogate density which maximizes both entropy and likelihood. It should be noted, however, that these respective maximizations are carried out over different domains. The mathematical content of Theorems (4.2) and (4.3) is little more than a restatement of the well-known result that in random samples from exponential populations one obtains the same estimate of the unknown population density by maximizing entropy and maximizing likelihood (see, for example, Dutta 1966; Campbell 1970).

Even though likelihood, as defined by (4.46), can be interpreted probabilistically in a random sampling context, its meaning here does not depend on underlying probability concepts. At this stage of our enquiry it is simply a function of data which turns out to be useful in data analysis and which suggests analogies with inferential statistics. Its phenomenological interpretation is examined in §9.

4.8. *Concluding remarks*

The preceding results show that classical statistical analysis based on flat exponential models can be interpreted in a purely descriptive way in terms of arity-restricted projective data condensation. Ideas like sufficiency, the maximization of entropy and the maximization of likelihood arise, not as *ad hoc* principles but, simply as a consequence of designating a surrogate density for use in an arity-restricted context to achieve consistency within the associated exponential family $e(\Phi_*)$. Theorem 4.2 shows that this family arises in a natural way when one maximizes entropy; but, by itself, this is not a compelling reason for restricting surrogation to the densities in that family. Nevertheless there are some important differences between the viewpoint adopted here and that of classical statistics. In the first place $\hat{f}_A$ is viewed as the designated surrogate for the suppressed data density; it is not seen, as it is in classical statistics, as an estimate of an unknown population density. In the second place, there is no explicit reference to a mechanism, stochastic or otherwise, which might have generated the data.

Seeing $\hat{f}_A$ as a surrogate for the suppressed data density f_A is closely related to the viewpoint in exploratory data analysis when one compares $\hat{f}_A$ with f_A, for instance by graphical procedures involving residuals, to ascertain how well the model fits the data. The probabilistic theory of statistical inference has to do with formulating such eye-ball assessments in a more precise way. These more precise assessments involve goodness-of-fit criteria and the sampling distributions of corresponding measures of the discrepancy between f_A and $\hat{f}_A$. Nevertheless the underlying concern with how well the model fits the data shows that the exploratory viewpoint has much in common with the one adopted here. There remains, of course, the question of whether one can assess the discrepancy between f_A and $\hat{f}_A$ in a meaningful way without recourse to sampling distributions and we address this question in §6.2.

The absence of probability concepts from the data condensation framework suggests the possibility of clarifying those concepts by means of data condensation. Indeed it could be argued that the concept of probability arises out of the way we organize the multitude of our perceptions of the world and that this consists, at least in part, in their condensation to what is salient. With that in mind, it does not seem altogether pointless to develop a theory of data condensation in a non-probabilistic framework and then investigate its bearing, if any on probability concepts. The preceding discussion of Bayes's postulate and binary sequences can be seen as first steps in that direction. However, arity restricted projective contexts seem, on the face of it, too rudimentary to sustain a wide degree of generality in such investigations and so we turn instead to two less ambitious but more immediate problems: (i) in an arity restricted context with data condensed by projection onto $\boldsymbol{\Phi}$, to what extent can one justify using only the surrogate spectra in $\mathscr{E}(\boldsymbol{\Phi})$, and (ii) can one extend the procedures developed for projective contexts to general macrostandard contexts?

5. Surrogation in general contexts

A surrogation procedure for the general ρ-context is defined to be an algorithm P that assigns to each data-set D in $\mathscr{D}$ a corresponding surrogate spectrum $\tilde{F}_D = P(D)$, for use in place of F_D when D is condensed at the macrolevel ρ, such that data-sets with the same spectrum are assigned the same surrogate. It is a mapping $P:\mathscr{D}\rightarrow\mathscr{G}$ with the property

$$F_{D'} = F_{D''} \Rightarrow P(D') = P(D''). \tag{5.1}$$

Among these mappings there are some which seem to be more useful than others. For instance, if we require consistency at the data-set A, then P must be such that equation (3.1) holds with $\tilde{F}_A = P(A)$. We now discuss two other properties, computability and equal informativeness.

5.1. *Computability*

A surrogation procedure P is said to be ρ-computable when

$$D'\rho D'' \Rightarrow P(D') = P(D''). \tag{5.2}$$

The assigned surrogate spectrum is then a function of the data condensation at macrolevel ρ and hence, in principle, computable from it. There is a sense in which one cannot do otherwise than use surrogation procedures that conform to (5.2), because if one needed something more than the data condensation to compute the surrogate spectrum, then that something would be, in effect, a contextual constraint which should be part of the contextual macrolevel. The viewpoint adopted here is that the data condensation $\rho(D)$ embodies everything that is known or supposed known about the data-set D. From that viewpoint ρ-computability is a statement about the meaning of surrogation at the macrolevel ρ.

More precisely, suppose one set out to use a surrogation procedure P at macrolevel ρ that was not ρ-computable. There would be data-sets D', D'' in $\mathscr{D}$ with $D'\rho D''$ but

$P(D') \neq P(D'')$. Let χ be the equivalence on $\mathscr{D}$ for which $E\chi D$ means $P(E) = P(D)$. Then $\tau = \chi \cap \rho$ is an equivalence on $\mathscr{D}$ with $\sigma \subseteq \tau \subset \rho$. The macrolevel τ is finer than ρ and it determines a corresponding macrostandard context at macrolevel τ. If δ represents ρ, then η given by $\eta(D) = (\delta(D), P(D))$ represents τ. Since η distinguishes between at least one pair of ρ-equivalent data-sets, e.g. D' and D'' mentioned above, we are working at a finer macrolevel than ρ. Moreover since $E\tau D$ implies that $P(E) = P(D)$, the surrogation procedure P is computable at the level τ at which we are then working. If for some D in $\mathscr{D}$ we required more than its condensation $\rho(D)$ to compute the surrogate spectrum $P(D)$, then this would discriminate between data sets at the macrolevel τ which is finer than the level ρ at which we claimed to be working, and call into question the correctness of saying that the macrolevel of the enquiry was ρ.

In what follows we adopt computability without further comment. All surrogation procedures are to be taken as computable at the macrolevels of the contexts in which they are used. In the context of section (4.2), the surrogation procedure $P(D) = N\hat{f}_D$ is computable with respect to the projection then in question.

5.2. *Equal informativeness*

Because of (5.2), a ρ-computable surrogation procedure assigns the same surrogate spectrum to each of the data-sets in a ρ-equivalence class. This common surrogate spectrum might supply different amounts of information about the various spectra for which it deputizes and this could be seen as inappropriate because those spectra are indistinguishable at the macrolevel ρ. To examine this possibility we adopt the usual logarithmic measure of information and define the amount of information provided by case n of the data-set $D = x_1 x_2 \dots x_N$ to be

$$I_n(D) = \ln F_D(x_n). \tag{5.3}$$

The total information in D is defined to be

$$I(D) = \sum_{n=1}^{N} I_n(D) = (F_D, \ln F_D). \tag{5.4}$$

When $\tilde{F}_D$ is a surrogate spectrum used in place of F_D, the substitutional surrogate version of (5.3) is

$$\tilde{I}_n(D) = \ln \tilde{F}_D(x_n), \tag{5.5}$$

whereas that of equation (5.4) is

$$I(D \,|\, \tilde{F}_D) = \sum_{n=1}^{N} \tilde{I}_n(D) = (F_D, \ln \tilde{F}_D). \tag{5.6}$$

This is the actual total of the individual pieces of surrogate information supplied by the cases in D when F_D is replaced by $\tilde{F}_D$. In general it is not computable from the condensation $\rho(D)$ alone, even when $\tilde{F}_D$ is derived from a ρ-computable surrogation procedure, because we need F_D to compute the inner product on the right of (5.6).

A surrogation procedure P is said to be equally informative about $\rho(D)$ when

$$I\{D' \,|\, P(D)\} = I\{D'' \,|\, P(D)\}, \forall\, D', D'' \in \rho(D), \tag{5.7}$$

or, equivalently, from (5.6), when

$$F_{D'} - F_{D''} \perp \ln P(D), \forall\, D', D'' \in \rho(D). \tag{5.8}$$

Such procedures can be characterized in the following way. Write

$$\nabla\{\rho(D)\} = \{F_{D'} - F_{D''} : D', D'' \in \rho(D)\} \tag{5.9}$$

for the set of differences of the spectra of the data-sets in $\rho(D)$ and let

$$\Phi\{\rho(D)\} = [\nabla\{\rho(D)\}]^{\perp} \tag{5.10}$$

be the subspace of $\mathscr{V}$ orthogonal to it. From (5.8) we obtain the following theorem.

Theorem 5.1. *The surrogation procedure P is equally informative about $\rho(D)$ if and only if the surrogate spectrum $P(D)$ belongs to the exponential family $\mathscr{E}[\Phi\{\rho(D)\}]$.*

It is advantageous to use surrogation procedures P which are simultaneously equally informative about each ρ-equivalence class, namely those for which

$$P(D) \in \mathscr{E}[\Phi\{\rho(D)\}], \forall D \in \mathscr{D}, \tag{5.11}$$

because they can be used whatever the data-set under study on any particular occasion. From a practical viewpoint it is parsimonious to work with a subset of these procedures. To do so write

$$\nabla(\rho) = \bigcup_{D \in \mathscr{D}} \nabla\{\rho(D)\}, \tag{5.12}$$

and

$$\Phi(\rho) = \{\nabla(\rho)\}^{\perp} = \bigcap_{D \in \mathscr{D}} \Phi\{\rho(D)\} \tag{5.13}$$

for the subspace of $\mathscr{V}$ orthogonal to it. The surrogation procedures P which are such that, for each A in $\mathscr{D}$, the surrogate spectrum $P(A)$ belongs to the exponential family

$$\mathscr{E}[\Phi(\rho)] = \bigcap_{D \in \mathscr{D}} \mathscr{E}[\Phi\{\rho(D)\}] \tag{5.14}$$

are simultaneously equally informative about each ρ-equivalence class. Such procedures are said to be fully ρ-informative. A surrogation procedure P is fully ρ-informative when

$$\forall D \in \mathscr{D} : \mathscr{P}(\mathscr{D}) = \exp(\phi_D), \quad \phi_D \in \Phi(\rho). \tag{5.15}$$

The elements of the exponential family $\mathscr{E}\{\Phi(\rho)\}$ are called the macrosurrogate spectra of the ρ-context. The surrogate spectra determined by a fully informative surrogation procedure are macrosurrogate spectra. The following result should be noted.

Theorem 5.2. *If ρ is the macrolevel corresponding to data condensation by projection onto Φ, then $\Phi(\rho) \supseteq \Phi$.*

Proof. Since $D'\rho D''$ means that $F_{D'} - F_{D''} \perp \Phi$, every ϕ in Φ belongs to each $\Phi\{\rho(D)\}$. Thus $\Phi \subseteq \Phi(\rho)$.

In particular, the restriction in §4 to the surrogate spectra in $\mathscr{E}(\Phi)$ is a restriction to macrosurrogate spectra.

Equally informative surrogation procedures could be questioned on the grounds that information about some data-sets might be more important than it is about others. For example, one might think of replacing (5.7) by

$$\frac{I\{D' \mid P(D)\}}{W(D')} = \frac{I\{D'' \mid P(D)\}}{W(D'')}, \forall D, D'' \in \rho(D), \tag{5.16}$$

where the $W(D)$ are positive weights which reflect the perceived relative importance of the data-sets in $\mathscr{D}$, the more important D the bigger its weight. Although this suggestion has some force, unequal weights in (5.16) would discriminate between the data-sets in a ρ-equivalence class contrary to the supposition that they are indistinguishable at the macrolevel ρ of the enquiry. As with violations of ρ-computability, we would then be working at a finer macrolevel than ρ and (5.7) would hold at that finer macrolevel. However, our viewpoint here is not that one must use equally informative surrogation procedures but rather the clarification of what is involved when we either do or not use them. In any event, as we will see later, there are other ways of quantifying the informational concept underlying $I\{D \mid \tilde{F}_D\}$ of (5.6).

5.3. *Projective closure*

Let π be the projection determined by the subspace $\boldsymbol{\Phi}$ of $\mathscr{V}$ through the logical equivalence (2.3). When $\boldsymbol{\Phi} = \{0\}$ contains only the 0-vector, π is the universal relation u on $\mathscr{D}$ for which any pair of data-sets of $\mathscr{D}$ are u-related. Thus any macrolevel ρ is contained in at least one projection, for example u. Suppose now that $\{\boldsymbol{\Phi}_t : t \in T\}$ is a non-empty family of subspaces of $\mathscr{V}$ and let π_t be the projective macrolevel corresponding to data condensation by projection onto $\boldsymbol{\Phi}_t$. Write

$$\boldsymbol{\Phi} = \bigvee_{t \in T} \boldsymbol{\Phi}_t \tag{5.17}$$

for the smallest subspace of $\mathscr{V}$ containing all the $\boldsymbol{\Phi}_t$ and

$$\pi = \bigcap_{t \in T} \pi_t \tag{5.18}$$

for the binary relation on $\mathscr{D}$ which is the set intersection of all the π_t. Then π is the projective macrolevel corresponding to data condensation by projection onto $\boldsymbol{\Phi}$. In other words, the set intersection of any non-empty family of projective macrolevels is itself a projective macrolevel. It follows that for each macrolevel ρ there is a smallest projective macrolevel which contains it, namely the set intersection of all the projective macrolevels which do contain it, that family being non-empty since the universal relation u belongs to it. We call the smallest projective macrolevel containing ρ the projective closure of ρ and denote it by $\bar{\rho}$. A projection is its own closure. The following theorem exhibits the structure of $\bar{\rho}$ in terms of that of ρ.

Theorem 5.3. *The projective closure $\bar{\rho}$ of the macrolevel ρ is the projective macrolevel corresponding to data condensation by projection onto the subspace $\boldsymbol{\Phi}(\rho)$ of* (5.13).

Proof. Let π be the macrolevel corresponding to projection onto $\boldsymbol{\Phi}(\rho)$. When $D'\rho D''$, $F_{D'} - F_{D''}$ is in $\nabla(\rho)$ of (5.12) and so

$$D'\rho D'' \Rightarrow F_{D'} - F_{D''} \perp \boldsymbol{\Phi}(\rho) \Rightarrow D'\pi D''.$$

In other words $\rho \subseteq \pi$. To show that π is the projective closure of ρ we verify that if $\lambda \supseteq \rho$ is any projection containing ρ, then $\pi \subseteq \lambda$. To do so suppose that λ corresponds to projection onto the subspace $\boldsymbol{\Psi}$ of $\mathscr{V}$ and write $\mu = \pi \cap \lambda$ for the projection determined by $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\rho) \vee \boldsymbol{\Psi}$, the subspace spanned by $\boldsymbol{\Phi}(\rho)$ and $\boldsymbol{\Psi}$. Since $\rho \subseteq \mu$, $D'\rho D''$ implies that $D'\mu D''$, and so, since

$$D'\mu D'' \Leftrightarrow F_{D'} - F_{D''} \perp \boldsymbol{\Phi},$$

every vector ϕ in $\boldsymbol{\Phi}$ is orthogonal to each difference $F_{D'} - F_{D''}$ with $D'\rho D''$. It follows that $\boldsymbol{\Phi} \subseteq \boldsymbol{\Phi}(\rho)$. Since $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\rho) \vee \boldsymbol{\Psi}$ we must have $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\rho)$ and $\boldsymbol{\Psi} \subseteq \boldsymbol{\Phi}(\rho)$. In other words, $\mu = \pi$ and $\pi \subseteq \lambda$.

The following result is worth noting.

Theorem 5.4. *If $\bar{\rho}$ is the projective closure of ρ, then*

$$\Phi(\bar{\rho}) = \Phi(\rho). \tag{5.19}$$

Proof. Since $\rho \subseteq \tau$ implies that $\nabla(\rho) \subseteq \nabla(\tau)$, we have

$$\rho \subseteq \tau \Rightarrow \Phi(\tau) \subseteq \Phi(\rho). \tag{5.20}$$

In particular $\Phi(\bar{\rho}) \subseteq \Phi(\rho)$. To establish the reverse inclusion, and hence (5.19), note that a vector $\theta \perp \Phi(\rho)$ is a finite linear combination of spectral differences $F_{D'} - F_{D''}$ with $D'\bar{\rho}D''$ and hence, by Theorem (5.3), with $F_{D'} - F_{D''} \perp \Phi(\rho)$. It follows that $\theta \perp \Phi(\rho)$ and hence that $\Phi(\rho) \subseteq \Phi(\bar{\rho})$.

Corollary. *The macrosurrogate spectra of the $\bar{\rho}$-context are the same as those of the ρ-context.*

Projective closure can also be characterized by means of the likelihood function. This is the content of the following theorem.

Theorem 5.5. *If $\bar{\rho}$ is the projective closure of ρ, then*

$$D'\bar{\rho}D'' \Leftrightarrow \operatorname{lik}(g \mid D') = \operatorname{lik}(g \mid D''), \forall g \in e\{\Phi(\rho)\}.$$

Proof. If $g = \mathrm{e}^{\phi}/(1, \mathrm{e}^{\phi})$, with ϕ in $\mathscr{E}\{\Phi(\rho)\}$, then (4.30) gives

$$\operatorname{lik}(g \mid D) = \exp[(\phi, F_D) - \ln(1, \mathrm{e}^{\phi})].$$

Thus
$$\operatorname{lik}(g \mid D')/\operatorname{lik}(g \mid D'') = \exp[(\phi, F_{D'} - F_{D''})],$$

and the theorem is an immediate consequence.

Corollary. *A macrolevel ρ is its own projective closure if and only if $\Phi(\rho)$ separates the ρ-equivalence classes in the sense that, when $\rho(D') \neq \rho(D'')$ there is a density g in $e\{\Phi(\rho)\}$ with* $\operatorname{lik}(g \mid D') \neq \operatorname{lik}(g \mid D'')$.

Proof. When $\Phi(\rho)$ separates the ρ-equivalence classes the forward implication in Theorem (5.5) shows that $\bar{\rho} \subseteq \rho$ and hence that $\rho = \bar{\rho}$. Conversely when $\bar{\rho} = \rho$, the backward implication shows that $\Phi(\rho)$ separates ρ-classes.

It is an immediate consequence of (5.2) and $\rho \subseteq \bar{\rho}$ that

Theorem 5.6. *A $\bar{\rho}$-computable surrogation procedure is ρ-computable.*

Finally we note this theorem.

Theorem 5.7. *If ρ is arity-restricted, that is ρ-equivalent data-sets have the same arity, then $\bar{\rho}$ is also arity restricted.*

Proof. For any data-sets D', D'' the inner product $(1, F_{D'} - F_{D''})$ is the difference between the arities of D', D''. Thus from (5.9), if ρ is arity restricted 1 is orthogonal to each $\nabla\{\rho(D)\}$, that is $\Phi(\rho)$ contains 1 and hence $\bar{\rho}$ is arity restricted. Note that the converse result is trivially true because $\rho \subseteq \bar{\rho}$.

In the next section we examine consistency in the $\bar{\rho}$-context and resolve the difficulty mentioned at the end of §3.

5.4. *Consistency and closure*

In a ρ-context, the macrolevel ρ contains the symmetry relation σ and so there is a function H such that

$$\forall D \in \mathscr{D} : \rho(D) = H(F_D). \tag{5.21}$$

Thus the representation-free version of the A-designation equation (3.1) is

$$H(\tilde{F}_A) = H(F_A). \tag{5.22}$$

Since $\rho \subseteq \bar{\rho}$, there is a function $\mathit{\Gamma}$ such that

$$\forall D \in \mathscr{D} : \bar{\rho}(D) = \mathit{\Gamma}\{\rho(D)\} = \mathit{\Gamma} \circ H(F_D) \tag{5.23}$$

and the representation-free version of the A-designation equation of the $\bar{\rho}$-context is

$$\mathit{\Gamma} \circ H(\tilde{F}_A) = \mathit{\Gamma} \circ H(F_A). \tag{5.24}$$

It follows at once that any surrogate spectrum $\tilde{F}_A$ satisfying (5.22) must also satisfy (5.24). In other words, any surrogate spectrum which is A-consistent for the ρ-context is also A-consistent for the associated $\bar{\rho}$-context.

Suppose that ρ is arity-restricted so that, by Theorem (5.7), $\bar{\rho}$ is also arity-restricted, let N be the arity of A and write $\boldsymbol{\Phi}_*(\rho)$ for $\boldsymbol{\Phi}(\rho) \cap 1^{\perp}$. By the result established in §4.3, the designation equation (5.24) of the $\bar{\rho}$-context has one and only one solution in $\mathscr{E}\{\boldsymbol{\Phi}(\rho)\}$, namely

$$\hat{F}_A = N\hat{f}_A, \tag{5.25}$$

where $\hat{f}_A$ is the unique density in the exponential family $e\{\boldsymbol{\Phi}_*(\rho)\}$ determined by the analogue of equation (4.6) for the $\bar{\rho}$-context. Thus, by virtue of the corollary to Theorem (5.4), there is at most one macrosurrogate spectrum of the ρ-context which is A-consistent and, when it exists, it is the unique A-consistent macrosurrogate spectrum of the $\bar{\rho}$-context. Thus we resolve the difficulty mentioned at the end of §3 by designating our surrogate spectra through consistency in the $\bar{\rho}$-context instead of the originating ρ-context. This avoids the difficulty of defining the function H of (5.21) on surrogate spectra because $\mathit{\Gamma} \circ H$ of (5.24) is well defined by the equation

$$\mathit{\Gamma} \circ H = \mathbb{P}_{\boldsymbol{\Phi}_*(\rho)}, \tag{5.26}$$

as projection onto $\boldsymbol{\Phi}_{**}(\rho)$. Moreover, with $H(\tilde{F}_A)$ defined in a way consistent with (5.26), if (5.22) did have a macrosurrogate solution it would have to be the one obtained by solving (5.24).

This way of designating a surrogate spectrum in an arity-restricted ρ-context is a ρ-computable surrogation procedure. For in the $\bar{\rho}$-context the surrogation procedure P which consists in taking $P(A)$ to be $\hat{F}_A$ of (5.25) is $\bar{\rho}$-computable, that is $P(D') = P(D'')$ when $D'\bar{\rho}D''$, because the macrolevel $\bar{\rho}$ corresponds to data condensation by a projection. Since $\rho \subseteq \bar{\rho}, D'\rho D''$ implies that $D'\bar{\rho}D''$ and hence that $P(D') = P(D'')$; in other words P is a ρ-computable surrogation procedure when it is applied to the originating ρ-context.

When ρ is the projective macrolevel defined by the condensation of data D to the projection of F_D onto $\boldsymbol{\Phi}$ we have $\boldsymbol{\Phi} \subseteq \boldsymbol{\Phi}(\rho)$, this was noted by Theorem (5.2). However, that theorem does not exclude the possibility that $\boldsymbol{\Phi}$ is a proper subset of $\boldsymbol{\Phi}(\rho)$. By theorem (5.3) $\bar{\rho}$, the projective closure of ρ, corresponds to data condensation by projection onto $\boldsymbol{\Phi}(\rho)$. But since ρ is a projective macrolevel, $\bar{\rho} = \rho$ and so

$$\mathbb{P}_{\boldsymbol{\Phi}} F_{D'} = \mathbb{P}_{\boldsymbol{\Phi}} F_{D''} \Leftrightarrow \mathbb{P}_{\boldsymbol{\Phi}(\rho)} F_{D'} = \mathbb{P}_{\boldsymbol{\Phi}(\rho)} F_{D''}$$

or equivalently,

$$F_{D'} - F_{D''} \perp \boldsymbol{\Phi} \Leftrightarrow F_{D'} - F_{D''} \perp \boldsymbol{\Phi}(\rho),$$

and, when $\boldsymbol{\Phi}$ is a proper subspace of $\boldsymbol{\Phi}(\rho)$, the projective macrolevel ρ can be generated by projection onto either of them. This possibility raises no difficulties from the viewpoint of surrogate designation because the designated surrogate density will then be in $e(\Phi_*)$. For the designation equation (4.6) has a unique solution in $e(\Phi_*)$ and, since $e(\Phi_*) \subset e\{\Phi_*(\rho)\}$ this must be the unique solution in $e\{\Phi_*(\rho)\}$ to the corresponding equation with Φ_* replaced by $\Phi_*(\rho)$.

5.5. *Concluding remarks*

The results obtained in this section can be seen as a satisfactory resolution of the issues raised at the end of §4. The use of the surrogate spectra in the exponential family $\mathscr{E}(\Phi)$ is related to the adoption of equal informativeness and parsimony in mathematical development, and while these are not compelling reasons they do show that their use is not entirely arbitrary. A deeper reason is discussed in §9.4. Secondly the procedures developed for projective contexts are extended to general macrostandard contexts by an appropriate use of projective closures.

There are two aspects of projective closure which are worth noting. The first is that the relationship between a macrolevel and its projective closure is similar to that in classical statistics between sufficiency and minimal sufficiency, but since this analogy is peripheral to our main concerns we do not examine it further here. The second is the relationship between projective closure and the linearization of a nonlinear data condensation. The projective macrolevel π of (2.3), corresponding to data condensation by projection onto the subspace Φ, can be represented by the vector-valued condensing statistic ϕ of equation (2.4). In that equation each component real-valued condensing statistic $\phi_m(D)$ is the linear aggregation of the function ϕ_m over the data-set which is given by equation (2.5). Data condensation by means of nonlinear aggregation over the data-set will, in general, generate a non-projective macrolevel. By working in the macrostandard context corresponding to its projective closure we transform the nonlinear data condensation into a linear one.

It should be noted that the results obtained so far lead to the designation of a uniquely determined macrosurrogate spectrum, which is computable from the data condensation, without any consideration of how well that surrogate fits the suppressed data spectrum. From the viewpoint adopted here the issue of goodness of fit is not a question of the best use of a given data condensation but the problem of what data condensation to use. In other words, it concerns the choice of macrolevel. Assessment of surrogate performance is examined in the next section. Because of the reduction to projective contexts by closure, we focus on projective contexts when it is convenient to do so.

6. Surrogate performance

The assessment of surrogate performance involves two related but separate issues (i) how well a proposed surrogate depicts the actual spectrum suppressed in the data condensation, and (ii) how effective similar condensations might be for other data-sets. Both issues are practically important but we are principally concerned here with only the first of them. The second is discussed briefly in §9.5. We focus on ordered pairs $[F_D, \tilde{F}_D]$ consisting of a data spectrum F_D and a possible surrogate for it, $\tilde{F}_D$. Such a pair is called a depiction and $[F_D, \tilde{F}_D]$ is said to be the depiction of F_D by $\tilde{F}_D$.

Throughout the section we suppose that we are dealing with an arity restricted context, A is the actual data-set being considered and N is its arity.

6.1. *Proximity to actuality*

We measure proximity between a data spectrum F_D and its surrogate $\tilde{F}_D$ by a scaled version of the difference between the information in D and the information about D contained in $\tilde{F}_D$. The amounts of information in question are taken to be $I(D)$ of (5.4) and $I(D\,|\,\tilde{F}_D)$ of (5.6) and the assessment of how well $\tilde{F}_D$ depicts F_D is based on the quantity

$$\varDelta[F_D, \tilde{F}_D] = N^{-1}[I(D) - I(D\,|\,\tilde{F}_D)], \tag{6.1}$$

the smaller its value the better the depiction of F_D by $\tilde{F}_D$. We call $\varDelta[F_D, \tilde{F}_D]$ the information deviance of the depiction $[F_D, \tilde{F}_D]$. Writing $F_D = Nf_D, F_D = N\tilde{f}_D$ we obtain the information deviance in terms of densities, namely

$$\varDelta[F_D, \tilde{F}_D] = (\ln f_D, f_D) - (\ln \tilde{f}_D, f_D) \tag{6.2}$$

and (4.41) shows that it is positive except when $\tilde{f}_D = f_D$. Equation (6.2) also shows that information deviance is the measure of nearness of probability distributions introduced by Kullback & Liebler (1951). We can also regard information deviance as a likelihood-ratio statistic because, from (4.46), $N(\ln g, f_A)$ is the log-likelihood $l_A(g)$ and so

$$\varDelta[F_D, \tilde{F}_D] = (1/N)\{l_D(f_D) - l_D(\tilde{f}_D)\} \tag{6.3}$$

$$= \frac{1}{N} \ln \left\{\frac{\text{lik }(f_D\,|\,D)}{\text{lik }(\tilde{f}_D\,|\,D)}\right\}. \tag{6.3}$$

In particular, the deviance of classical statistics is $2N$ times information deviance.

It follows from (6.3) that if $\tilde{F}'_D$ and $\tilde{F}''_D$ are two possible surrogates for F_D, then

$$\varDelta[F_D, \tilde{F}'_D] \leqslant \varDelta[F_D, \tilde{F}''_D) \Leftrightarrow \text{lik }(\tilde{f}'_D\,|\,D) \geqslant \text{lik }(\tilde{f}''_D\,|\,D). \tag{6.4}$$

In other words the bigger the D-likelihood of $\tilde{F}_D$, the better the depiction of F_D by $\tilde{F}_D$ as measured by information deviance.

If ρ is the projective macrolevel corresponding to data condensation by projection onto $\boldsymbol{\varPhi}$, then, by Theorem (4.2), the unique density $\hat{f}_A$ in $e(\boldsymbol{\varPhi}_*)$ which is $\boldsymbol{\varPhi}_*$-consistent at A maximizes A-likelihood over $e(\boldsymbol{\varPhi}_*)$. In other words, recalling the closing remarks of §5.4, the designated macrosurrogate density $\hat{f}_A$ leads to the best depiction of F_A by a macrosurrogate spectrum of the ρ-context. From (4.48) and (6.3), the information deviance of that depiction is the difference between the entropies of the designated macrosurrogate density, and the suppressed actual density. Introducing the dependence on the macrolevel explicitly, by writing $\hat{F}_{A|\rho}$ and $\hat{f}_{A|\rho}$ for the designated macrosurrogate spectrum and density at the projective macrolevel ρ, we obtain

$$\varDelta[F_A, \hat{F}_{A|\rho}] = \text{Ent }(\hat{f}_{A|\rho}) - \text{Ent }(f_A). \tag{6.5}$$

The following theorem shows that the finer the projective macrolevel, the better the depiction.

Theorem 6.1. *If $\rho \subseteq \tau$ are arity-restricted macrolevels, then*

$$\varDelta[F_A, \hat{F}_{A|\bar{\rho}}] \leqslant \varDelta[F_A, \hat{F}_A|_{\bar{\tau}}].$$

Proof. Both $\bar{\rho}$ and $\bar{\tau}$ are arity-restricted by Theorem (5.7). By the implication (5.20), we have $\boldsymbol{\varPhi}(\tau) \subseteq \boldsymbol{\varPhi}(\rho)$ and hence $e\{\boldsymbol{\varPhi}_*(\tau)\} \subseteq e\{\boldsymbol{\varPhi}_*(\rho)\}$. Thus both $\hat{f}_{A|\bar{\rho}}$ and $\hat{f}_{A|\bar{\tau}}$ belong to $e\{\boldsymbol{\varPhi}_*(\rho)\}$. But by theorem (4.2) $\hat{f}_{A|\bar{\rho}}$ maximizes A-likelihood over $e\{\boldsymbol{\varPhi}_*(\rho)\}$, hence

$$\text{lik }(\hat{f}_{A|\bar{\tau}}\,|\,A) \leqslant \text{lik }(\hat{f}_{A|\bar{\rho}}\,|\,A)$$

and the theorem is an immediate consequence of equation (6.3).

The smallest value of $\varDelta[F_A, F_{A|\rho}]$ is 0. This arises when $\bar{\rho} = \sigma$, the symmetry equivalence, and corresponds to data condensation by projection onto $\mathscr{V}$. The context is then identifiable and $\hat{F}_{A|\sigma} = F_A$.

The largest value of $\varDelta[F_A, \tilde{F}_{A|\bar{\tau}}]$ arises when $\bar{\tau} = \alpha$ the arity equivalence, namely the macrolevel α for which $D'\alpha D''$ means that D', D'' have the same arity. This corresponds to data condensation by projection onto the one-dimensional subspace generated by the constant vector 1. For any other arity-restricted macrolevel ρ we have $\bar{\rho} \subseteq \alpha$ and so, by Theorem (6.1), the largest information deviance occurs when $\bar{\tau} = \alpha$. But, as in §4.4,

$$\hat{f}_{A|\alpha}(x) \equiv 1/V, x \in S, \tag{6.6}$$

with entropy

$$\text{Ent}\ (\hat{f}_{A|\alpha}) = \ln V, \tag{6.7}$$

and so, by (6.5),

$$\varDelta[F_A, \hat{F}_{A|\alpha}] = \ln V - \text{Ent}\ (f_A) \tag{6.8}$$

is the largest value of $\varDelta[F_A, \tilde{F}_{A|\bar{\tau}}]$.

6.2. *Assessing proximity*

Of two depictions of F_A, the one with the smaller information deviance is the better depiction. It is parsimonious to compress all possible pairwise comparisons into a smaller number of standardized comparisons. We may do so by adopting the worst case as a benchmark and, in the $\bar{\rho}$-context, calculating the quantity

$$\chi_A(\bar{\rho}) = 100\left[\frac{\varDelta[F_A, \hat{F}_{A|\alpha}] - \varDelta[F_A, \hat{F}_{A|\bar{\rho}}]}{\varDelta[F_A, \hat{F}_{A|\alpha}]}\right]. \tag{6.9}$$

This is the percentage deviance reduction achieved by condensing the data-set A at the macrolevel $\bar{\rho}$ instead of at the arity macrolevel α. By Theorem (6.1), it increases with increasing fineness of resolution in the macrolevel $\bar{\rho}$, from 0% at $\bar{\rho} = \alpha$ to 100% at $\bar{\rho} = \sigma$. From (6.5) and (6.8),

$$\chi_A(\bar{\rho}) = 100\left[\frac{\ln V - \text{Ent}\ (\hat{f}_{A|\bar{\rho}})}{\ln V - \text{Ent}\ (f_A)}\right]. \tag{6.10}$$

Using (4.47) and (4.48) we obtain the equivalent expression

$$\chi_A(\bar{\rho}) = 100\left[\frac{N \ln V + l_A(\hat{f}_{A|\bar{\rho}})}{N \ln V + l_A(f_A)}\right]. \tag{6.11}$$

Percentage deviance reduction assesses how well $\hat{F}_{A|\bar{\rho}}$, the designated macro-surrogate spectrum at the macrolevels ρ or $\bar{\rho}$, depicts the suppressed data spectrum F_A by comparing it to data condensation by arity alone. To compute it we need to calculate the entropy, or equivalently, the likelihood of the suppressed data density. When the percentage deviance reduction is small we are led to investigate the possibility of getting a better depiction by data condensation at a finer macrolevel with correspondingly higher percentage deviance reduction and smaller information deviance. In the next section we illustrate its use by a numerical example.

6.3. *A numerical example*

Consider the dose-response mortality data of Bliss (1935) which have often been used for illustrative purposes. For our purposes it is convenient to discuss it within the framework of the explanatory surrogation leading to the logistic regression

Table 1. *Beetle mortality data* (*Bliss 1935*)

			conditional death rates	
dose ($\log_{10}(CS_2/(\text{mg l}^{-1}))$) ξ	number of insects $F_A(\xi,\cdot)$	number dead $F_A(\xi,1)$	actual $f_A(1\mid\xi)$	surrogate $\hat{f}_A(1\mid\xi)$
1.6907	59	6	0.102	0.119 (0.106)
1.7342	60	13	0.217	0.175 (0.196)
1.7552	62	18	0.290	0.306 (0.297)
1.7842	56	28	0.500	0.535 (0.530)
1.8113	63	52	0.825	0.781 (0.780)
1.8369	59	53	0.898	0.928 (0.928)
1.8610	62	61	0.984	0.981 (0.981)
1.8839	60	60	1.000	1.000 1.000
totals	481	291		

(4.35). The data are presented in the first three columns of table 1. There are 481 cases, 291 deaths and the profiles are the eight doses of the first column. The data condensation corresponds to the choices

$$\alpha_u(\xi) = \xi^{u-1}, \quad u = 1, 2, \ldots,$$
$$\beta_i(\eta) \equiv 0,$$

so that the condensation of the marginal spectrum $F_A(\cdot, \eta)$ is absent, and

$$\psi_j(\xi) = \delta(\xi, p_j), \quad 1 \leqslant j \leqslant J,$$

where $p_1, p_2, \ldots, p_J$ are the distinct profiles in the data and δ is a Kronecker delta, so that the condensation of $F_A(\xi, \cdot)$ is $F_A(\xi, \cdot)$ itself. Finally η is 1 or 0 according as the case is a death or a survivor.

We consider the four explanatory variables 1, ξ, ξ^2 and ξ^3 in the three combinations (i) 1 and ξ, (ii) 1, ξ and ξ^2, (iii) 1, ξ, ξ^2 and ξ^3. When we use only 1 and ξ, the data is condensed to the number of cases column, namely the one headed $F_A(\xi, \cdot)$, together with the total number of deaths and the total lethal profile $\Sigma\, \xi F_A(\xi, 1)$. In the two other explanatory combinations we successively add into the data condensation the additional total lethal profiles $\Sigma\, \xi^2 F_A(\xi, 1)$ and $\Sigma\, \xi^3 F_A(\xi, 1)$. The surrogate conditional death rates in table 1 are the $\hat{f}_A(1\mid\xi)$ when $1, \xi$ and ξ^2 are used as explanatory variables. The associated numbers in parentheses will be discussed later.

The use of 1 and ξ alone as explanatory variables gives 95.4% deviance reduction. This rises to 98.7% when we add ξ^2 to the explanatory set. This confirms the impression gained from a comparison of the actual conditional rates in table 1 with their surrogates based on 1, ξ and ξ^2, namely that the data condensation in question is quite effective. The classical deviance is then 3.19, slightly less than the 3.45 deviance of Dobson's (1983) extreme value model. The use of ξ^3 as an additional explanatory variable does not lead to an appreciable improvement in the percentage deviance reduction, it rises by only 0.1–98.8%. These calculations we performed using (6.10) with $V = 15$, not 16, because there are no data survivors at the highest dosage.

6.4. *The likelihood principle*

We can compute a percentage deviance reduction

$$\chi[F_A,\tilde{F}_A] = 100\left[\frac{\Delta[F_A,\hat{F}_{A|\alpha}]-\Delta[F_A,\tilde{F}_A]}{\Delta[F_A,\hat{F}_{A|\alpha}]}\right] \tag{6.12}$$

for any depiction $[F_A,\tilde{F}_A]$ of F_A by a possible surrogate $\tilde{F}_A$. As at (6.11) we have

$$\chi[F_A,\tilde{F}_A] = 100\left[\frac{N\ln V+l_A(\tilde{f}_A)}{N\ln V+l_A(f_A)}\right]. \tag{6.13}$$

From (6.3) we can obtain the information deviance $\Delta[F_A,\tilde{F}_A]$ for any surrogate spectrum $\tilde{F}_A$ once we know the A-likelihood function lik $(\cdot\,|\,A)$. The same is true of the percentage deviance reduction (6.13). Thus we have the following theorem.

Theorem 6.2. *To assess how well a surrogate spectrum depicts the actual spectrum of the data-set A by means of information deviance it is enough to know the A-likelihood function.*

This theorem can be seen as a descriptive version of the likelihood principle. But it does not endorse a general principle always favouring greater likelihoods, even when these come from different data-sets. On the contrary the reverse is true for data-sets within a macroequivalence class. For while $\hat{F}_{A|\bar{\rho}}$ is the best depiction of F_A by a macrosurrogate spectrum of the ρ-context, whatever the actual data-set A, just how good that depiction is will depend on which of the data sets in the condensing macrodatum $\rho(D)$ is the one at hand. Indeed, since $D\rho E$ implies that $\hat{F}_{D|\bar{\rho}} = \hat{F}_{E|\bar{\rho}}$ it follows from (6.3) that when $D\rho E$

$$l_D(f_D) \geqslant l_E(f_E) \Leftrightarrow \Delta[F_D,\hat{F}_{D|\bar{\rho}}] \geqslant \Delta[F_E,\hat{F}_{E|\bar{\rho}}]. \tag{6.14}$$

In other words, the greater the D-likelihood of f_D as D varies within a ρ-equivalence class, the *worse* the depiction of F_D by $\hat{F}_{D|\bar{\rho}}$. The data-sets best condensed at level ρ are the ones with the least likelihood, and hence the largest entropy, within the ρ-equivalence class to which they belong. The disparity between the consequences of increasing likelihood in (6.4) and (6.14) arises because in (6.4) we consider the A-likelihood lik $(f|A)$ with A fixed but f varying, whereas in (6.14) we are involved with the D-likelihood lik $(f_D\,|\,D)$ with D varying within a ρ-equivalence class. The fact the data-sets best condensed at level ρ are the ones with the least likelihood within a ρ-class suggests that there might be some advantage in taking ρ to be the macrolevel for which the likelihood function is constant within a ρ-class. The data-set A would then be condensed to its arity and the common value of the likelihood in the class to which A belongs, namely lik $(f_A\,|\,A)$. But this leads to an identifiable context. This follows from the fact that the inequality (4.45) is an equality only when $g = f_A$. Thus there is then only one density corresponding to the ρ-class $\rho(A)$, namely f_A, and so F_A is, in principle, recoverable from lik $(f_A\,|\,A)$ and the arity of A.

6.5. *Concluding remarks*

At the beginning of this section we noted that assessment of surrogate performance involves not only how well we depict the suppressed data spectrum but also how effective similar condensations might be for other data-sets. In classical statistics it is difficult to draw a sharp line between these two aspects of data analysis, because they are usually treated together. Proximity between the data density and the modelling density, namely the estimated population density, is seen as indicating a goodness-of-fit which tells us not only that the modelling density depicts the data correspondingly well, but also that a similar modelling procedure will be effective, and produce essentially the same results, for data-sets which are typical random samples from the population then in question.

Inferential statistics focuses on the sampling distribution of the information deviance $\Delta[F_A, F_{A|\rho}]$ when A is regarded as a random sample from a population with a given density f. Improbably large values of the deviance are seen as discrediting the hypothesis that the population density is f. This procedure is particularly useful in the rebuttal of an ill-considered claim that data exhibits some especially noteworthy features. For if typical random samples from a population without those features exhibit them to an extent like that in the data, then chance rather than something substantive might well be seen as a satisfactory explanation of them. The inferential outlook is, in part, an anticipatory calculation aimed at showing that such a rebuttal of what we claim would lack substance. But its emphasis on an underlying population makes it difficult to address the data *per se*, without reference to mechanisms which might have generated it. Shifting the emphasis from an estimate of an unknown population density to a surrogate for the suppressed data density makes it easier to distinguish between questions of inference on the one hand and those of description on the other. For instance, the question of how close $\hat{F}_{A|\rho}$ is to F_A, which has been the principal concern in this section, is a question about the condensation of A at macrolevel ρ and, though we might also be interested in how our answer to it changes with A, it is not itself a question about random sampling from a population.

Although per cent reduction in deviance is a useful indicator, not only of how well $\hat{F}_{A|\rho}$ depicts F_A, but also of the improvements gained and the losses incurred when we condense A at other macrolevels, it does not answer all of the questions that might be asked about a particular data condensation. For instance, even if the percentage deviance reduction is high, we might be interested in the possibility of doing better. If there are relatively few better depictions, then we might settle for what we have already obtained. If there are relatively many more effective depictions then we might think it worthwhile to search for one which better suits our purposes. Thus we might also be interested in the distribution of $\Delta[F_A, \hat{F}_{A|\bar{\tau}}]$ as $\bar{\tau}$ runs through all the arity restricted projective macrolevels and determining, in some appropriate limiting sense, the proportion of $\bar{\tau} \subseteq \alpha$ for which $\Delta[F_A, \hat{F}_{A|\bar{\tau}}] \leqslant \Delta[F_A, \hat{F}_{A|\bar{\rho}}]$. Similarly one might be interested in the distribution of $\Delta[F_A, \hat{F}_{A|\bar{\rho}}]$ as A runs through a contextually relevant family of data-sets. These interests would lead one to the descriptive methods of Finch (1981); we do not pursue them here.

Instead, we take up the possibility raised at the end of §5.2 and examine another way of quantifying the amount of information about D which is contained in the surrogate spectrum $\tilde{F}_D$. This will lead us to consider data condensation by the projection of $\ln F_D$ onto a subspace $\boldsymbol{\Phi}$ as a possible alternative to data condensation by the projection of F_D onto $\boldsymbol{\Phi}$. This sort of data condensation has a number of attractive features which stem from the fact that the designation equation

$$\mathbb{P} \ln \tilde{F}_A = \mathbb{P} \ln F_A \tag{6.15}$$

has the unique solution

$$\check{F}_A = \exp\{\mathbb{P} \ln F_A\} \tag{6.16}$$

in $\mathscr{E}(\Phi)$. This surrogate spectrum is easy to compute and, perhaps more importantly, since it has an explicit closed form, issues relating to how $\check{F}_A$ changes with changes in A and Φ are much easier to investigate than are the corresponding issues for $\hat{F}_A$. Moreover in the classical random sampling framework the associated densities $\check{f}_A$ and $\hat{f}_A$ are asymptotically equivalent. In particular the surrogate density $\check{f}_A$ can be interpreted as a large sample approximation to the ML-estimate in a flat exponential family. From our descriptive viewpoint, however, the two types of data condensation arise from different ways of quantifying total information.

7. Metric information

In this section we examine the consequences of measuring total information in terms of the metric structure of the space $\mathscr{V}$ instead of by the simple additions of equations (5.4) and (5.6). We retain the definition (5.3) but now think of the function $\ln F_D(x)$ as an information vector in $\mathscr{V}$ and define the metric information in D to be

$$J(D) = \| \ln F_D \|, \tag{7.1}$$

namely the norm in $\mathscr{V}$ of the information vector $\ln F_D$. The metric information about D in the surrogate spectrum $\tilde{F}_D$ is now defined to be the magnitude of the component of the information vector $\ln F_D$ in the direction of its surrogate $\ln \tilde{F}_D$, namely

$$J(D \,|\, \tilde{F}_D) = J(D) \cos \{\omega(F_D, \tilde{F}_D)\}, \tag{7.2}$$

where $\cos\{\omega(F_D, \tilde{F}_D)\}$, the cosine of the angle between $\ln F_D$ and $\ln \tilde{F}_D$, is non-negative because each of those vectors is strictly positive on the support S. From (7.1) and (7.2) we obtain

$$J(D \,|\, \tilde{F}_D) = (\ln F_D, \ln \tilde{F}_D)/\| \ln \tilde{F}_D \|. \tag{7.3}$$

7.1. *Equal metric informativeness*

As at (5.7), a surrogation procedure P is said to be equally metric-informative about $\rho(D)$ when

$$J\{D' \,|\, P(D)\} = J\{D'' \,|\, P(D)\}, \forall\, D', D'' \in \rho(D),$$

or, from (7.3), equivalently when

$$\ln F_{D'} - \ln F_{D''} \perp \ln P(D), \forall\, D', D'' \in < \rho(D). \tag{7.4}$$

Such procedures can be characterized by results which are analogues of those in §5.2. Write

$$A\{\rho(D)\} = \{\ln F_{D'} - \ln F_{D''} : D', D'' \in \rho(D)\} \tag{7.5}$$

for the set of differences of the natural logarithms of the spectra of the data-sets in $\rho(D)$ and write

$$\Phi^0\{\rho(D)\} = [A\{\rho(D)\}]^{\perp} \tag{7.6}$$

for the subspace orthogonal to it. Then we have

Theorem 7.1. *The surrogation procedure P is equally metric-informative about $\rho(D)$ if and only if the surrogate spectrum $P(D)$ belongs to the exponential family $\mathscr{E}[\Phi^0\{\rho(D)\}]$.*

The surrogation procedures P which are simultaneously equally metric-informative about each ρ-class are those for which

$$P(D) \in \mathscr{E}[\Phi^0\{\rho(D)\}], \forall\, D \in \mathscr{D}.$$

Write
$$\Lambda(\rho) = \bigcup_{D\in\mathscr{D}} \Lambda\{\rho(D)\} \tag{7.7}$$

and
$$\varPhi^0(\rho) = \{\Lambda(\rho)\}^{\perp} = \bigcap_{D\in\mathscr{D}} \varPhi^0\{\rho(D)\}. \tag{7.8}$$

The surrogation procedures P such that, for each Λ in $\mathscr{D}$, the surrogate spectrum $P(\Lambda)$ belongs to the exponential family
$$\mathscr{E}\{\varPhi^0(\rho)\} = \bigcap_{D\in\mathscr{D}} \mathscr{E}[\varPhi^0\{\rho(D)\}] \tag{7.9}$$

are simultaneously equally metric informative about each ρ-equivalence class. Such procedures are said to fully metric ρ-informative. A surrogation procedure P is fully metric ρ-informative when
$$\forall D\in\mathscr{D}: P(D) = \exp\,(\phi_D), \quad \phi_D \in \varPhi^0(\rho). \tag{7.10}$$

The elements of the exponential family $\mathscr{E}\{\varPhi^0(\rho)\}$ are called the metric-surrogate spectra of the ρ-context. As in §5.2 it is parsimonious to use the fully metric ρ-informative surrogation procedures, and hence only metric-surrogate spectra, when one wants to ensure equally metric informativeness.

7.2. *Prologjections*

Let $\varPhi$ be a subspace of $\mathscr{V}$ and define the macrolevel λ on $\mathscr{D}$ by the logical equivalence
$$D'\lambda D'' \Leftrightarrow \ln F_{D'} - \ln F_{D''} \perp \varPhi. \tag{7.11}$$

This macrolevel can be represented by the condensing statistic
$$\eta(D) = \mathbb{P}\ln F_D. \tag{7.12}$$

It can also be represented in terms of linearly independent vectors $\phi_0, \phi_1, \ldots, \phi_M$ spanning $\varPhi$ by means of the vector-valued condensing statistic
$$\phi^0(D) = (\phi^0_0(D), \phi^0_1(D), \ldots, \phi^0_M(D)), \tag{7.13}$$

where
$$\phi^0_m(D) = (\phi_m, \ln F_D), \quad 0 \leqslant m \leqslant M. \tag{7.14}$$

The condensing statistic η of (7.12) is called a prologjection and the macrolevel λ of (7.11) is said to be prologjective. The following analogue of Theorem (5.2) follows from the fact that each ϕ in $\varPhi$ belongs to each $\varPhi^0\{\lambda(D)\}$ of (7.6)

Theorem 7.2. *If λ is the prologjective macrolevel corresponding to data condensation by prologjection onto $\varPhi$, then $\varPhi^0(\lambda) \supseteq \varPhi$.*

When $\varPhi = \{0\}$ any two data-sets in $\mathscr{D}$ are λ-equivalent. When $\varPhi$ contains the constant vector 1,
$$(\ln F_D, 1) = \sum_S \ln F_D(x) \tag{7.15}$$

is constant within each λ-equivalence class; we call $(\ln F_D, 1)$ the larity of D. When $\varPhi$ is the subspace of constant vectors we denote the corresponding prologjective macrolevel by α^0 and call it the larity macrolevel. Thus $D'\alpha^0 D''$ means that D', D'' have the same larity. An arbitrary prologjection is said to be larity-restricted when its $\varPhi$ contains 1, that is when $\eta \subseteq \alpha^0$. When $\varPhi = \mathscr{V}$ the prologjection is the symmetry macrolevel σ.

7.3. *Prologjective closure*

The arguments of §5.3 can be extended in an obvious way to prologjections. The smallest prologjective macrolevel containing the macrolevel ρ is called the prologjective closure of ρ and denoted by ρ^0. A prologjection is its own prologjective closure. The following theorem is the analogue of Theorem 5.3. It is proved by replacing F_D by $\ln F_D$ in the proof of that theorem.

Theorem 7.3. *The prologjective closure ρ^0 of the macrolevel ρ is the prologjective macrolevel corresponding to data condensation by prologjection onto the subspace $\Phi^0(\rho)$ of* (7.8).

Similarly we have the following.

Theorem 7.4. *If ρ^0 is the prologjective closure of the macrolevel ρ, then*

$$\Phi^0(\rho^0) = \Phi^0(\rho). \tag{7.16}$$

Corollary. *The metric-surrogate spectra of the ρ^0-context are the same as those of the ρ-context.*

Theorem 7.5. *A ρ^0-computable surrogation procedure is ρ-computable.*

Theorem 7.6. *If ρ is larity-restricted, that is ρ-equivalent data-sets have the same larity, then ρ^0 is also larity-restricted.*

7.4. *Consistency in prologjective contexts*

Let λ be the prologjective macrolevel (7.11) which corresponds to data condensation by prologjection onto the subspace Φ. The A-designation equation of the λ-context in the η-representation of (7.12) is

$$\mathbb{P} \ln \tilde{F}_A = \mathbb{P} \ln F_A. \tag{7.17}$$

Its solution is given by the next theorem.

Theorem 7.7. *There is a unique surrogate spectrum of the form $\tilde{F}_A = \exp(\phi_A)$ with ϕ_A in Φ which satisfies the A-designation equation of the λ-context. It is*

$$\check{F}_A = \exp(\mathbb{P} \ln F_A). \tag{7.18}$$

Proof. The surrogate spectrum $\check{F}_A$ of (7.18) satisfies (7.17). Conversely if $\tilde{F}_A = \exp(\phi_A)$ with ϕ_A in Φ does satisfy (7.18), then

$$\phi_A = \mathbb{P}\phi_A = \mathbb{P} \ln \tilde{F}_A = \mathbb{P} \ln F_A$$

and so $$\tilde{F}_A = \check{F}_A.$$

Since λ is its own prologjective closure, it can also be represented by prologjection onto the subspace $\Phi^0(\lambda)$. By Theorem 7.7, the A-designation equation of the λ-context in that representation has a unique solution $\tilde{F}^0_A = \exp(\phi^0_A)$ with Φ^0_A in $\phi^0(\lambda)$. But since $\Phi \subseteq \Phi^0(\lambda), \check{F}_A$ of (7.18) does have that form and so the element ϕ^0_A of $\Phi^0(\lambda)$ then the question is in fact $\mathbb{P} \ln F_A$ in Φ. In other words:

Theorem 7.8. *There is a unique metric-surrogate spectrum of the λ-context satisfying the A-designation equation of that context, it is $\check{F}_A$ of* (7.18).

Corollary 1. *If $\phi_0, \phi_1, \ldots, \phi_M$ is an orthonormal basis of $\varPhi$, then*

$$\check{F}_A(x) = \exp\left\{\sum_{m=0}^{M} (\ln F_A, \phi_m)\, \phi_m(x)\right\}, \quad x \in S. \tag{7.19}$$

Corollary 2. *The unique metric-surrogate spectrum of the λ-context minimizes*

$$\|\ln \tilde{F}_A - \ln F_A\|^2 \tag{7.20}$$

over the set of all the metric-surrogate spectra of that context.

Equation (7.19) gives the designated metric-surrogate spectrum directly in terms of the orthonormal condensing statistics $\phi^0_m(A)$ of equation (7.14). The following theorem is an immediate consequence of (7.18).

Theorem 7.9. *Let $\varPhi_1 \perp \varPhi_2$ be two orthogonal subspaces of $\mathscr{V}$ and let $\varPhi_3 = \varPhi_1 \vee \varPhi_2$ be the smallest subspace containing both of them. Let $\mathbb{P}_k$ be projection onto $\varPhi_k$ and let*

$$\check{F}_{kA} = \exp(\mathbb{P}_k \ln F_A)$$

be the designated metric-surrogate spectrum corresponding to data condensation by prologjection onto $\varPhi_k$. Then

$$\ln \check{F}_{3A} = \ln \check{F}_{1A} + \ln \check{F}_{2A}. \tag{7.21}$$

Using (7.21) it is easy to examine not only the effect of adding in further orthogonal condensing statistics to a given prologjective data condensation but also the contributions from specified subsets of orthogonal condensing statistics. In particular, in the context of Corollary 1 to Theorem 7.8 we have

$$\ln \check{F}_A = \sum_{m=0}^{M} \ln \check{F}_{mA}, \quad \ln \check{F}_{mA} = (\ln F_A, \phi_m)\, \phi_m. \tag{7.22}$$

This gives the individual contributions to $\check{F}_A$ from the component condensing statistics $\phi^0_m(A)$. In practice they can be presented in an informative and easily understood way by means of the condensing display matrix

$$[(\ln F_A, \phi_m)\, \phi_m(x)] \tag{7.23}$$

with $M+1$ rows and V columns. When V is large a graph of the rows serves the same purpose and highlights the ways in which the various condensing statistics affect different parts of the designated metric-surrogate spectrum.

7.5. *Metric-surrogation versus macrosurrogation*

The following theorem exhibits the relationship between surrogate designation in projective contexts and surrogate designation in prologjective contexts.

Theorem 7.10. *Let λ be the prologjective macrolevel corresponding to data condensation by prologjection onto the subspace $\varPhi$ and let ρ be the projective macrolevel corresponding to data condensation by projection onto the same subspace $\varPhi$. If $\check{F}_A$ is the designated metric-surrogate spectrum of the λ-context and $\hat{F}_A$ is the designated macrosurrogate spectrum of the ρ-context, then both $\check{F}_A$ and $\hat{F}_A$ are in $\mathscr{E}(\varPhi)$ and*

$$\check{F}_A = \hat{F}_A \exp\{-\mathbb{P} \ln (\hat{F}_A / F_A)\}. \tag{7.24}$$

Corollary.

$$\check{F}_A = F_A \exp\{\mathbb{P}^{\perp} \ln (\hat{F}_A / F_A)\}, \tag{7.25}$$

where $\mathbb{P}^{\perp}$ is projection onto $\varPhi^{\perp}$, the orthogonal complement of $\varPhi$.

Proof. Both $\ln \hat{F}_A$ and $\ln \check{F}_A$ are in $\varPhi$. Thus

$$\begin{aligned}\ln \check{F}_A - \ln \hat{F}_A &= \mathbb{P} \ln \check{F}_A - \mathbb{P} \ln \hat{F}_A \\ &= \mathbb{P} \ln F_A - \mathbb{P} \ln \hat{F}_A \\ &= -\mathbb{P}\,(\ln \hat{F}_A - \ln F_A).\end{aligned}$$

This is (7.24). To obtain (7.25) we rewrite (7.24) in the form

$$\ln (\check{F}_A/F_A) = \ln (\hat{F}_A/F_A) - \mathbb{P} \ln (\hat{F}_A/F_A).$$

A related result is the following theorem.

Theorem 7.11. *If F is any surrogate spectrum in $\mathscr{E}(\varPhi)$, then*

$$\check{F}_A = F \exp \{\mathbb{P} \ln F_A/F\}. \tag{7.26}$$

Proof. We start from

$$\ln F_A = \ln F + \ln (F_A/F).$$

Since $\ln F$ is in $\varPhi$ and $\ln \check{F}_A = \mathbb{P} \ln F_A$,

$$\ln \check{F}_A = \ln F + \mathbb{P} \ln (F_A/F).$$

This is (7.26).

We use this theorem to establish Theorem 7.12.

Theorem 7.12. *Let f be a density in $e(\varPhi)$ and let*

$$\check{f}_A = \check{F}_A/(1, \check{F}_A) \tag{7.27}$$

be the designated metric-surrogate density, then

$$\check{f}_A = NL^{-1} f \exp \{\mathbb{P} \ln (f_A/f)\}, \tag{7.28}$$

where N is the arity of A, $L = (1, \check{F}_A)$ is its surrogate larity and

$$NL^{-1} = (1, f \exp \{\mathbb{P} \ln (f_A/f)\}). \tag{7.29}$$

Proof. Put $F = Nf$ *so that* F_A/F is f_A/f. Equations (7.28) and (7.29) follow at once from (7.26).

The classical consistency of $\check{f}_A$ when it is regarded as an estimate of an underlying population density in $e(\varPhi)$ is a simple consequence of this theorem. For if sampling is random, sample size increases indefinitely and the population density f is in $e(\varPhi)$, then f_A converges to f with probability one. Equation (7.28) shows that under the same circumstances $\check{f}_A$ converges to f with probability one. If the population density f is not in $e(\varPhi)$, then we have the wrong model and f in (7.28) is replaced by $\exp (\mathbb{P} \ln f)$. In this case $\check{f}_A$ converges with probability one to $\exp (\mathbb{P} \ln f)/(1, \exp (\mathbb{P} \ln f))$. Equation (7.29) with $f = \hat{f}_A$ gives density analogues of (7.24) and (7.25), namely

$$\check{f}_A = NL^{-1} \hat{f}_A \exp \{-\mathbb{P} \ln (\hat{f}_A/f_A)\}, \tag{7.30}$$

and

$$\check{f}_A = NL^{-1} f_A \exp \{\mathbb{P}^{\perp} \ln (\hat{f}_A/f_A)\}. \tag{7.31}$$

It follows that $\check{f}_A$ is asymptotically equivalent to $\hat{f}_A$ is a classical setting, because $\check{f}_A/\hat{f}_A$ is close to 1 when that is true of $\hat{f}_A/f_A$.

To illustrate some of these results we return to the beetle mortality data in table 1 of §6.3. The numbers in parentheses in that table are the surrogate death rates $\check{f}_A(1 \mid \xi)$ derived from $\check{F}_A$ of (7.18) when $\varPhi$ is the subspace used in the calculation of the corresponding macrosurrogate death rates. The proximity of these $\check{f}_A(1\,|\,\xi)$ to the corresponding $\hat{f}_A(1\,|\,\xi)$ is transparent. It is interesting that in seven out of the eight profiles in table 1, $\check{f}_A(1\,|\,\xi)$ is closer to $f_A(1\,|\,\xi)$ than is $\hat{f}_A(1\,|\,\xi)$. To compare $\check{F}_A$ with $\hat{F}_A$ in this case we can compute the percentage deviance reduction of the depiction $[F_A, \check{F}_A]$ as given by (6.13). For the data of table 1, this is 98.2%, close to the 98.7% deviance reduction of the depiction $[F_A, \hat{F}_A]$.

From the descriptive viewpoint adopted here metric-surrogation has as much claim for consideration as does macrosurrogation. The general reasons for using either of them are much the same and differ only in the way we quantify the concept of total information. Although macrosurrogation leads to interesting descriptive interpretations of some of the important concepts of classical statistics, metric-surrogation has the practical advantages mentioned at the end of §6. Moreover metric surrogation is easily extended to include data-sets A whose spectral supports S_A are proper subsets of the common support S of the surrogate spectra in $\mathscr{E}(\Phi)$. To do so one need only replace the designation equation (7.17) by

$$\mathbb{P} \ln \tilde{F}_A = \mathbb{P}(I_A \ln F_A), \tag{7.32}$$

where I_A is the indicator function of S_A and, on the right, $0 \ln 0$ is interpreted as 0. This equation has the unique solution

$$\check{F}_A = \exp \mathbb{P}(I_A \ln F_A) \tag{7.33}$$

in $\mathscr{E}(\Phi)$.

The results obtained in the earlier subsections of this section are straightforward mathematical analogues of the corresponding results in §5. But their correctness as mathematical theorems does not in itself imply that it is equally correct to use them in data analysis. We see in §9 that there are reasons for preferring macrosurrogation to metric-surrogation. Nevertheless consistent metric-surrogation retains an independent interest as a possibly large data-set approximation to consistent macrosurrogation.

8. Circumstantial condensation

In the preceding sections it was supposed that data condensation arose as a deliberate attempt to capture the gist of the data in a relatively simple way. There are also situations in which data condensation arises by force of circumstance, for instance when analysing the data involves censoring and truncation. We call this circumstantial condensation and illustrate it by examining surrogation in the analysis of survival data.

8.1. *Survival data*

Consider a data-set

$$A = x_1, x_2, \ldots, x_N, \tag{8.1}$$

in which each case-reading $x_n = (\eta_n, \tau_n)$ is an ordered pair of non-negative integers that arises in the following way. The cases are patients who enter a clinical trial at various times and are followed until they develop an end-point that is certain to occur sometime and which we will call 'death'. All times are measured in days and there is a common time origin day 0 for all the patients. Case n enters the trial on day τ_n and dies η_n days later; the possibility that $\eta_n = 0$ is not excluded.

In practice such a clinical trial is analysed before all the patients have died and, perhaps, before all of them have entered into it. If the data is analysed on day T, then one has no reading for a case with $\tau_n > T$ and the cases with $\tau_n \leqslant T$ fall into two categories. For the then deaths, namely the cases with $\eta_n + \tau_n \leqslant T$, one has the full case-reading $x_n = (\eta_n, \tau_n)$, but for the then survivors, namely the cases for which $\eta_n > T - \tau_n \geqslant 0$, one has only τ_n and the fact that η_n is bigger than $T - \tau_n$. For simplicity we consider only times T which exceed the entry times of all the patients under consideration. The censoring of lifetimes at time T can be viewed as a projective data condensation in the following way.

The support S_A of the data-set A is a subset of the non-negative quadrant of the (η, τ)-plane. At time T the points in S_A which are on or below the line $\eta + \tau = T$ correspond to the then deaths and the points in S_A which are above that line correspond to the then survivors. Censoring at time T partitions the support S_A into one-point cells on or below the line $\eta + \tau = T$ and into cells above that line which are horizontal strips at the entry times of the survivors. The effect of the censoring is to condense the data to the number of cases in each cell of the partition. In the next section we show that this sort of data condensation corresponds to projection of the data spectrum onto the subspace generated by the indicator functions of the cells of the partition.

For ease of comparison with standard survival analysis it is convenient to work with the censoring times $\zeta_n = T - \tau_n$ and to regard the data-set at time T as given by

$$A = y_1, y_2, \ldots, y_N, \tag{8.2}$$

where the reading for case n is

$$y_n = (\eta_n, \zeta_n), \quad 1 \leqslant n \leqslant N. \tag{8.3}$$

The deaths and survivors correspond to $\eta_n \leqslant \zeta_n$ and $\eta_n > \zeta_n$ respectively. The support at time T is then a subset of the non-negative quadrant of the (η, ζ)-plane and the censoring partitions it into one-point cells on or above the line $\eta = \zeta$ and into cells below that line which are horizontal strips at the censoring times of the survivors. In §§8.4 and 8.5 we consider the more general situation in which (8.3) is replaced by

$$r_n = (\xi_n, \eta_n, \zeta_n), \quad 1 \leqslant n \leqslant N, \tag{8.4}$$

where ξ_n is a possibly vector-valued explanatory case-profile.

Survival data present additional complications which stem from the nature of the data support. At time T, when the data set is given by (8.2), the data support is the set of distinct (η, ζ) in the data but these points are not all known at that time, because of the censoring. We know that the points (η, ζ) corresponding to the deaths are in the data-set, but for a survivor with censoring time ζ we know only that the corresponding lifetime η is a positive integer exceeding ζ. Thus we are dealing with a macrostandard context in which the data support is not known. This calls for a modification of the procedures developed earlier. The simplest modification, and the one we adopt here, is to remove the restriction that the designated surrogate spectrum should have the same support as the data spectrum. To do so we consider surrogate spectra with a given support S which contains the data support S_A. In a projective context involving data condensation by projection of F_A onto a vector space Φ of functions on S this means that we look for solutions $\tilde{F}_A$ in $\mathscr{E}(\Phi)$ to the designation equations (3.3) when the data spectrum F_A is no longer strictly positive on S.

We suppose that the first patients enter the trial on day 1 so that $0 < \tau_n < T$ and $1 \leqslant \zeta < T$. Since a lifetime ending on or before day T can be at most $T-1$ days long, we truncate lifetimes by T. In other words, we replace η_n in (8.3) by $\min(\eta_n, T)$. Finally the underlying support S is taken to be the data pairs (η, ζ) with $\eta \leqslant \zeta$, together with the points in the sets

$$\Gamma(+, \zeta) = \{(\zeta+1, \zeta), (\zeta+2, \zeta), \dots, (T, \zeta)\} \tag{8.5}$$

for each $\zeta, 1 \leqslant \zeta < T$, which is the censoring time of at least one survivor.

Accounts of clinical trials sometimes report only the lifetimes of the deaths and the censoring times of the survivors. This corresponds to a partition of S by the horizontal strips $\Gamma(+, \zeta)$ of (8.5) together with the vertical strips

$$\Gamma(\eta, *) = \{(\eta, \eta), (\eta, \eta+1), \dots, (\eta, T)\} \tag{8.6}$$

at each η which is the lifetime of at least one death. We see below that this does not affect the surrogate lifetime density. However, such accounts often fail to report the value of T and in those circumstances it is convenient to take T to be its smallest possible value. In other words, if μ is the maximum of the $\min(\eta_n, \zeta_n)$, then we take $T = \mu+1$ if there is a survivor with its $\zeta = \mu$ and take $T = \mu$ otherwise.

8.2. *Condensation by support partitioning*

Suppose that the generic support S is partitioned into M non-empty mutually disjoint cells $\Gamma_1, \Gamma_2, \dots, \Gamma_M$ which have union S and respective sizes $\gamma_1, \gamma_2, \dots, \gamma_M$. Let ϕ_m be the indicator function of the cell Γ_m, they are mutually orthogonal and $\|\phi_m\|^2 = \gamma_m$, let Φ be the subspace of V which they generate.

Consider the condensation of a data set A with support $S_A \subseteq S$ by the projection of its spectrum F_A onto Φ. In the representation determined by the orthogonal basis vectors $\phi_1, \phi_2, \dots, \phi_M$, the condensing statistic

$$\phi_m(A) = (\phi_m, F_A) = \sum_{\Gamma_m} F_A(x) = F_A(\Gamma_m) \tag{8.7}$$

is the number of data readings in the cell Γ_m. Thus data condensation by the projection of F_A onto Φ corresponds to grouping the data by the cells of the partition and recording only the multiplicities of the readings in them.

The surrogate spectra in $\mathscr{E}(\Phi)$ have the form $G(x) = \exp \phi(x)$ with

$$\phi = \sum_{m=1}^{M} (\phi, \phi_m) \ \gamma_m^{-1} \phi_m \tag{8.8}$$

constant within each cell of the partition. If such a spectrum satisfies the designation equations $(\phi_m, G) = (\phi_m, F_A)$, $1 \leqslant m \leqslant M$, then

$$F_A(\Gamma_m) = (\phi_m, G) = \gamma_m \exp \{(\phi, \phi_m) \gamma_m^{-1}\}, \quad 1 \leqslant m \leqslant M. \tag{8.9}$$

If each cell of the partition is represented in the data set, then each $F_A(\Gamma_m)$ is positive and these equations determine the (ϕ, ϕ_m) and hence, from (8.8), the vector ϕ. In this case there is a unique solution from $\mathscr{E}(\Phi)$ to the designation equations. It is constant within cells and is given in Γ_m by

$$\hat{F}_A(x) = \gamma_m^{-1} F_A(\Gamma_m), \quad x \in \Gamma_m. \tag{8.10}$$

The corresponding density is

$$\hat{f}_A(x) = \gamma_m^{-1} f_A(\Gamma_m), \quad x \in \Gamma_m, \tag{8.11}$$

where $f_A(\Gamma_m)$ is the data density of readings in Γ_m. In other words, in each cell of the partition the designated macrosurrogate spectrum and density are the arithmetic means of the corresponding data quantities in that cell. This type of data condensation was discussed in *ad hoc* ways in Finch (1977, 1980*a*, *b*, 1982*a*).

With $\hat{F}_A$ given by (8.10) we have $(\phi_m, \hat{F}_A) = F_A(\mathit{\Gamma}_m)$ even if $F_A(\mathit{\Gamma}_m)$ is 0. Thus $\hat{F}_A(x)$ of (8.10) is a solution to the designation equations even when some of the cells are not represented in the data. But in that case its support is the union of the cells that are represented in the data and it does not belong to the exponential family $\mathscr{E}(\boldsymbol{\mathit{\Phi}})$, except in a limiting sense at infinity. This may be seen from (8.9) which gives $(\phi, \phi_m) = -\infty$ when $F_A(\mathit{\Gamma}_m) = 0$.

One could also condense the data by prologjection onto $\boldsymbol{\mathit{\Phi}}$. The condensing statistics would then be

$$\phi_m^{\circ}(A) = (\phi_m, I_A \ln F_A) = \ln\left[\prod_{\mathit{\Gamma}_m} \{F_A(x)\}^{I_A(x)}\right], \tag{8.12}$$

where I_A is the indicator function of S_A the support of A. A short calculation shows that the designated metric-surrogate spectrum is also constant within cells and is given in $\mathit{\Gamma}_m$ by

$$\check{F}_A(x) = \left[\prod_{\mathit{\Gamma}_m} \{F_A(x)\}^{I_A(x)}\right]^{1/\gamma_m}, \quad x \in \mathit{\Gamma}_m, \tag{8.13}$$

namely by the geometric mean of the data spectrum in that cell. The corresponding density is

$$\check{f}_A(x) = \left[\prod_{\mathit{\Gamma}_m} \{f_A(x)\}^{I_A(x)}\right]^{1/\gamma_m}, \quad x \in \mathit{\Gamma}_m. \tag{8.14}$$

The metric-surrogate density $\check{F}_A$ of (8.12) is in $\mathscr{E}(\boldsymbol{\mathit{\Phi}})$ and has support S even when some cells of the partition are not represented in the data. For if the cell $\mathit{\Gamma}_m$ is not represented in the data then (8.13) gives $\check{F}_A(x) = 1$ because $0^{\circ} = \exp\,(0 \ln 0)$ is to be interpreted as 1. Metric surrogation is not available in the context of survival data because the censoring produces the condensing statistics (8.7), not those of (8.12).

The deviance reduction associated with $\hat{F}_A$ of (8.10) is

$$\mathit{\Delta}[F_A, \hat{F}_{A|\alpha}] - \mathit{\Delta}[F_A, \hat{F}_A] = \ln V - \mathrm{Ent}\,(\hat{f}_A), \tag{8.15}$$

where
$$V = \sum_{m=1}^{M} \gamma_m \tag{8.16}$$

is the size of S and

$$\mathrm{Ent}\,(\hat{f}_A) = \sum_{m=1}^{M} f_A(\mathit{\Gamma}_m)\{\ln \gamma_m - \ln f_A(\mathit{\Gamma}_m)\}. \tag{8.17}$$

When the data condensation is circumstantial we cannot compute the percentage deviance reduction of (6.10) because the actual density f_A is not then known. In such cases the deviance reduction associated with a partition is informative when one wants to assess the gain in moving from one given partition to a finer one. We illustrate this in §8.4.

8.3. *Surrogate survival*

We now apply the results just obtained to the partition of the subset S of the (η, ζ)-plane associated with the censoring of lifetimes by the analysis of survival data at time T. The underlying support S is partitioned into the one-point data pairs (η, ζ)

with $\eta \leqslant \zeta$ and the sets $\Gamma(+,\zeta)$ of (8.5), and each cell of this partition is represented in the data. From (8.11) the designated macrosurrogate density is $\hat{f}$ given on S by

$$\left.\begin{aligned}\hat{f}(\eta,\zeta) &= f(\eta,\zeta), && \eta \leqslant \zeta,\\ &= f(+,\zeta)/(T-\zeta), && \eta > \zeta,\end{aligned}\right\} \tag{8.18}$$

where f is the data density and $f(+,\zeta)$ is the proportion of cases in the data that are survivors with censoring time ζ. Writing $f(\eta,*)$ for the proportion of cases in the data that are deaths at lifetime η, namely the sum of the $f(\eta,\zeta)$ over the ζ with (η,ζ) in S and $\zeta \geqslant \eta$, equation (8.18) shows that the surrogate marginal density for η is

$$\hat{f}(\eta,\cdot) = f(\eta,*) + \sum_{\zeta<\eta} \frac{f(+,\zeta)}{(T-\zeta)}, \tag{8.19}$$

where η is the lifetime of at least one point in S and summation is over those ζ with (η,ζ) in S and $\zeta < \eta$. If η is not the lifetime of at least one point in the support S, then $\hat{f}(\eta,\cdot)$ is 0. From (8.19) we can obtain the corresponding surrogate survival function

$$\hat{S}(\eta) = \sum_{\eta',\,\eta'\geqslant\eta} \hat{f}(\eta',\cdot). \tag{8.20}$$

This is essentially the flat survival function of Finch (1977), where the underlying support S was the rectangle of all the integer pairs (η,ζ) with $0 \leqslant \eta \leqslant T$ and $1 \leqslant \zeta < T$.

If S is partitioned both by the $\Gamma(+,\zeta)$ of (8.5) and the $\Gamma(\eta,*)$ of (8.6), then the first part of equation (8.18) is replaced by

$$\hat{f}(\eta,\zeta) = f(\eta,*)/(T-\eta+1), \quad \eta \leqslant \zeta. \tag{8.21}$$

In this case we obtain the same surrogate marginal lifetime density as before, namely (8.19).

If there are no survivors, then (8.18) gives $\hat{f} = f$ whereas (8.21) spreads $f(\eta,*)$ uniformly across the cell $\Gamma(\eta,*)$. In both cases $\hat{f}(\eta,\cdot) = f(\eta,\cdot)$.

In figure 1, the surrogate survival function (8.20) is compared to the Kaplan–Meier survival function for the treatment group of the data considered in the next section. As that figure shows, surrogate survival along the tail is less optimistic than the corresponding Kaplan–Meier estimate. The Kaplan–Meier survival function would arise as a surrogate survival function if one partitioned S by the $\Gamma(+,\zeta)$ of (8.5) and the $\Gamma(\eta,*)$ of (8.6) and used a consistent surrogate density g with the independence property

$$g(\eta,\zeta) = g(\eta,\cdot)\,g(\cdot,\zeta) \tag{8.22}$$

on S.

To see this let $Y(t)$ be the subset of S consisting of the (η,ζ) in S with $\eta \geqslant t$ and let $Z(t)$ be the set of the (η,ζ) in S with $\zeta \geqslant t$. Let $R(t)$ be the set intersection of $Y(t)$ and $Z(t)$. Writing $g(B)$ for the sum of the $g(\eta,\zeta)$ over a subset B of S, equation (8.22) gives

$$g\{R(t)\} = g\{Y(t)\}\,g\{Z(t)\}. \tag{8.23}$$

Noting that the set of (η,ζ) in S with $\eta \geqslant t+1$ and $\zeta \geqslant t$ is $R(t)-D(t)$ with

$$D(t) = \{Y(t)-Y(t+1)\} \cap Z(t), \tag{8.24}$$

equation (8.22) also gives

$$g\{R(t)-D(t)\} = g\{Y(t+1)\}\,g\{Z(t)\}. \tag{8.25}$$

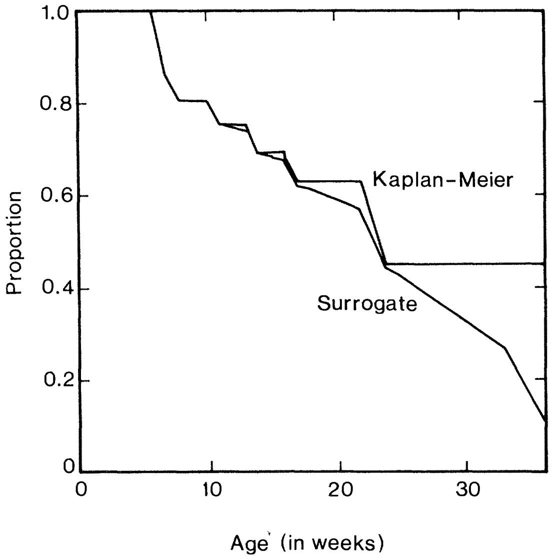

Figure 1. Surrogate and Kaplan–Meier survival functions for the treatment group of table 2.

If $g\{R(t)\} \neq 0$, then equations (8.23) and (8.25) give

$$g\{Y(t+1)\} = [1-g\{D(t)\}/g\{R(t)\}]\,g\{Y(t)\}. \tag{8.26}$$

But

$$g\{R(t)\} = \sum_{\eta \geqslant t} g(\eta, *) + \sum_{\zeta \geqslant t} g(+, \zeta)$$

and if g is consistent, then

$$g(\eta, *) = f(\eta, *), \quad g(+, \zeta) = f(+, \zeta),$$

and hence

$$g\{R(t)\} = f\{R(t)\} \tag{8.27}$$

can be calculated from the data condensation, namely the multiplicities in the cells of the partition. Moreover $D(t)$ of (8.24) is the set of data deaths at lifetime t and so

$$g\{D(t)\} = g(t, *) = f(t, *) = f\{D(t)\}. \tag{8.28}$$

Thus equation (8.26) gives

$$g\{Y(t+1)\} = [1-f\{D(t)\}/f\{R(t)\}]\,g\{Y(t)\}. \tag{8.29}$$

This is the well-known recurrence relation for the Kaplan–Meier survival function and so, under the independence constraint (8.22), the surrogate survival function is the Kaplan–Meier survival function.

From (8.23) we obtain the $g\{Z(t)\}$ and hence the individual $g(\eta, \zeta)$ from (8.22). This bivariate density may vary within the cells of the partition and so it is not, in general, a macrosurrogate density for the projective context based on those cells. It is, however, a macrosurrogate density for the macrostandard context derived from that projective context by adding to it the independence condition (8.22) as a contextual constraint. This would be the appropriate context when it is known that the actual data density f itself has the independence property (8.22). In that case, the argument leading to (8.29) from (8.22), with g replaced by f, shows that the Kaplan–Meier estimate of the survival function is then the data survival function.

It is worth noting that any consistent surrogate density $g(\eta,\zeta)$ leads to a product formula like (8.29) for its associated survival function, whether or not it has the independence property (8.22). For writing

$$\gamma(\eta,\zeta)=g(\eta,\zeta)-g(\eta,\cdot)\,g(\cdot,\zeta)$$

a straightforward argument shows that the analogue of equation (8.29) is

$$g\{Y(t+1)\}=\left[1-\frac{f\{D(t)\}-\gamma\{D(t)\}}{f\{R(t)\}-\gamma\{R(t)\}}\right]g\{Y(t)\} \tag{8.30}$$

for those t with $f\{R(t)\}>\gamma\{R(t)\}$. In (8.30) we always have $f\{R(t)\}\geqslant\gamma\{R(t)\}$ because

$$f\{R(t)\}-\gamma\{R(t)\}=g\{R(t)\}-\gamma\{R(t)\}=g\{Y(t)\}\,g\{Z(t)\}\geqslant 0.$$

Similarly we always have $f\{D(t)\}\geqslant\gamma\{D(t)\}$. In like manner, writing

$$\epsilon(\eta,\zeta)=f(\eta,\zeta)-f(\eta,\cdot)f(\cdot,\zeta)$$

we obtain
$$f\{Y(t+1)\}=\left[1-\frac{f\{D(t)\}-\epsilon\{D(t)\}}{f\{R(t)\}-\epsilon\{R(t)\}}\right]f\{Y(t)\}. \tag{8.31}$$

Since $f\{D(t)\}-\epsilon\{D(t)\}$ is $f(t,\cdot)\,f\{Z(t)\}$ and $f\{R(t)\}-\epsilon\{R(t)\}$ is $f\{Y(t)\}\,f\{Z(t)\}$, equation (8.31) simply states that $f\{Y(t+1)\}$ is $f\{Y(t)\}$ minus $f(t,\cdot)$. It does, however, display the way in which the data survival function differs from its Kaplan–Meier counterpart in the absence of independence, namely when the $\epsilon\{D(t)\}$ and the $\epsilon\{R(t)\}$ are not all zero.

8.4. *Comparison of surrogate survival in two groups*

Suppose now that there are two groups of patients, $G(0)$ and $G(1)$, and that correspondingly the readings take the form (8.4) where ξ is a binary 0, 1 variable specifying group membership, and allow for the possibility that T has different values for the two groups, $T(0)$ in $G(0)$ and $T(1)$ in $G(1)$. The support S consists of two sheets of the (η,ζ)-plane, one sheet for each group. Suppose it is partitioned by the cells

$$\left.\begin{aligned}\mathit{\Gamma}(\xi,+,\zeta)&=\{(\xi,\eta,\zeta):\eta=\zeta+1,\zeta+2,\ldots,T(\xi)\},\\ \mathit{\Gamma}(\xi,\eta,*)&=\{(\xi,\eta,\zeta):\zeta=\eta,\eta+1,\ldots,T(\xi)\}.\end{aligned}\right\} \tag{8.32}$$

The data is condensed to the number of cases in each cell of the partition; namely for each group, to the number of survivors and the number of deaths at each of the censoring times and lifetimes in question. From (8.11), the designated macrosurrogate density is $\hat{f}$ given on S by

$$\hat{f}(\xi,\eta,\zeta)=\begin{cases}f(\xi,\eta,*)/\{T(\xi)-\eta+1\}, & \eta\leqslant\zeta,\\ f(\xi,+,\zeta)/\{T(\xi)-\zeta\}, & \eta>\zeta.\end{cases} \tag{8.33}$$

Thus the surrogate bivariate density for group membership and lifetime is

$$\hat{f}(\xi,\eta,\cdot)=f(\xi,\eta,*)+\sum_{\zeta>\eta}\frac{f(\xi,+,\zeta)}{\{T(\xi)-\zeta\}}. \tag{8.34}$$

Since $\hat{f}(\xi,\eta,*)=f(\xi,\eta,*)$ and $\hat{f}(\xi,+,\zeta)$ is $f(\xi,+,\zeta)$, we have $\hat{f}(\xi,\cdot,\cdot)=f(\xi,\cdot,\cdot)$ and so the surrogate conditional density for lifetimes given group membership is

$$\hat{f}(\eta\,|\,\xi)=\hat{f}(\xi,\eta,\cdot)/f(\xi,\cdot,\cdot). \tag{8.35}$$

Table 2. *Times in weeks of remission of leukemia patients* (*Cox 1972*)

drug 6-MP	6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35.
control	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.

In practice the two groups usually arise from two different treatments and it is pertinent whether survival under one of the treatments is better than it is under the other. In classical statistics this question is formulated as a test of the hypothesis of no difference between the survival experience in two underlying population groups. In the descriptive framework of this paper, the question has to do with how much better one does by surrogation based on the two-sheet partition of (8.32) than by ignoring group membership and basing it instead on the one-sheet partition of (8.5) and (8.6). It may be investigated by comparing the deviance reductions associated with the two partitions, namely the quantities given by (8.15). By way of illustration we consider the well-known survival data in table 2. For that data the one-sheet partition achieves a deviance reduction of 0.215 whereas that of the two-sheet partition is 0.522, about 2.4 times as great a reduction in the deviance. This suggests that group membership plays a useful role in analysing the data, namely that the data is appreciably better described by calculating different surrogate survival functions for the two groups than by one such function for the combined group. It should be noted that the issue addressed in this analysis is whether the actual patients in the trial had different survival experiences according to the treatment they received; not whether the two groups came from corresponding populations with different survival experience.

8.5. *Surrogate survival with profile averaging*

In this section we consider surrogate survival with an explanatory case-profile as at (8.4). The data-set has the form $A = r_1 r_2 \ldots r_N$ where r_n, the reading for case n, is an ordered triple (ξ_n, η_n, ζ_n) with ξ_n its profile, η_n its truncated lifetime, $0 \leqslant \eta_n \leqslant T$, and ζ_n its censoring time, $1 \leqslant \zeta_n < T$. Let X be the set of distinct profiles in the data. The generic support S is taken to be

$$S = \{(x, y, z) : x \in X, 0 \leqslant y \leqslant T, 1 \leqslant z < T\}, \tag{8.36}$$

where the y and z in question are integers, and it is partitioned by the sets

$$\left.\begin{aligned} \Gamma(y, *) &= \{(x, y, z) : x \in X \mathbin{\&} x \geqslant z\}, \\ \Gamma(+, z) &= \{(x, y, z) : x \in X \mathbin{\&} y > z\}. \end{aligned}\right\} \tag{8.37}$$

Data readings in $\Gamma(y, *)$ correspond to deaths with lifetime y and those in $\Gamma(+, z)$ correspond to survivors with censoring time z. The profiles in X are typically real vectors

$$x = (x_1, x_2, \ldots, x_K) \tag{8.38}$$

of variables that might explain case differences in survival and, correspondingly, interest focuses on the dependence of lifetime on explanatory profile as displayed by $\hat{f}(\eta \mid t(\xi))$, the surrogate conditional density for lifetime given a function of the profile. A detailed case-study of this sort of survival analysis and its comparison with more familiar procedures will be presented elsewhere. Here we illustrate the general method by condensing the data to the cell multiplicities association with the partition (8.37) and certain profile averages.

To construct the data condensation let ϕ_y^d and ϕ_z^s be the indicator functions of $\Gamma(y,*)$ and $\Gamma(+,z)$ respectively. Let π_u, $u=0,1,\ldots,U$ be linearly independent functions on X and put

$$\psi_u^d(\xi,\eta,\zeta)=\pi_u(\xi)\,H(\eta,\zeta),\quad \psi_u^s(\xi,\eta,\zeta)=\pi_u(\xi)\,\{1-H(\eta,\zeta)\}, \tag{8.39}$$

where $H(y,z)$ is 1 when $y\leqslant z$ and 0 otherwise. Let $\varPhi$ be the subspace spanned by the ϕ_y^d, ϕ_z^s, ψ_u^d and ψ_u^s and condense the data by projecting its spectrum onto $\varPhi$. The associated real-valued condensing statistics are (i) the number of data deaths at each support lifetime,

$$(\phi_y^d,F_A)=F_A\{\Gamma(y,*)\},\quad 0\leqslant y\leqslant T, \tag{8.40}$$

and (ii) the number of data survivors at each support censoring time,

$$(\phi_z^s,F_A)=F_A\{\Gamma(+,z)\},\quad 1\leqslant z<T, \tag{8.41}$$

together with (iii) the aggregate lethal profiles

$$(\psi_u^d,F_A)=\sum_{x,\,y\leqslant z}\pi_u(x)\,F_A(x,y,z),\quad 0\leqslant u\leqslant U, \tag{8.42}$$

and (iv) the aggregate survival profiles

$$(\psi_u^s,F_A)=\sum_{x,\,y>z}\pi_u(x)\,F_A(x,y,z),\quad 0\leqslant u\leqslant U. \tag{8.43}$$

Writing $$\Gamma(*)=\bigcup_{0\leqslant y\leqslant T}\Gamma(y,*),\Gamma(+)=\bigcup_{1\leqslant z<T}\Gamma(+,z),$$

the number of data deaths,

$$F_A\{\Gamma(*)\}=\sum_0^T F_A\{\Gamma(y,*)\},$$

is obtained from the condensing statistics (8.40), and the number of data survivors,

$$F_A\{\Gamma(+)\}=\sum_1^{T-1}F_A\{\Gamma(+,z)\}$$

is obtained from the condensing statistics (8.41). Thus $(\psi_u^d,F_A)/F_A\{\Gamma(*)\}$ is the average value of $\pi_u(x)$ over the profiles of the data deaths and $(\psi_u^s,F_A)/F_A\{\Gamma(+)\}$ is the average value of $\pi_u(x)$ over the profiles of the data survivors.

For example if the profiles are given by (8.38) and $\pi_k(x)=x_k$, then $(\psi_k^d,F_A)/F_A\{\Gamma(*)\}$ and $(\psi_k^s,F_A)/F_A\{\Gamma(+)\}$ are the means of the variable x_k over deaths and survivors. Similarly if $\pi_u(x)=x_i x_j$, then the condensing statistics (8.42) and (8.43) give the corresponding death and survival correlations of the explanatory variables x_i and x_j.

In the projective context described above the designated macrosurrogate spectrum is given on S by an expression of the form

$$\hat{F}(\xi,\eta,\zeta)=\exp\left[\Sigma\{b_y^d\phi_y^d+b_z^s\phi_z^s+c_u^d\psi_u^d+c_u^s\psi_u^s\}\right],$$

where the constants b_y^d, b_z^s, c_u^d and c_u^s are determined by the designation equations, namely the four equations for the condensing statistic with $\hat{F}$ replacing F_A on their left-hand sides. Because of the special forms of the vectors ϕ_y^d, etc., we have

$$\begin{aligned}\hat{F}(\xi,\eta,\zeta) &= \exp\{b_\eta^d + \sum_u c_u^d\,\pi_u(\xi)\}, \quad \eta \leqslant \zeta,\\ &= \exp\{b_\zeta^{\mathrm{s}} + \sum_u c_u^{\mathrm{s}}\,\pi_u(\xi)\}, \quad \eta > \zeta.\end{aligned}$$

The constants b_η^d and b_ζ^{s} can be eliminated by using the first two designation equations, namely

$$\exp\,(b_\eta^d)\sum_x \exp\,\{\sum_u c_u^d\,\pi_u(x)\} = F_A\{\Gamma(\eta,*)\}/(T-\eta+1),$$

$$\exp\,(b_\zeta^d)\sum_x \exp\,\{\sum_u c_u^{\mathrm{s}}\,\pi_u(x)\} = F_A\{\Gamma(+,\cdot)\}/(T-\zeta).$$

Writing $$\lambda(\xi) = \sum_u c_u^d\,\pi_u(\xi), \quad \sigma(\xi) = \sum_u c_u^{\mathrm{s}}\,\pi_u(\xi), \tag{8.44}$$

and using densities instead of spectra, we obtain

$$\left.\begin{aligned}\hat{f}(\xi,\eta,\zeta) &= \frac{f_A\{\Gamma(\eta,*)\}}{(T-\eta+1)}\,\frac{\mathrm{e}^{\lambda(\xi)}}{(1,\mathrm{e}^{\lambda})}, \quad \eta \leqslant \zeta,\\ &= \frac{f_A\{\Gamma(+,\zeta)\}}{(T-\zeta)}\,\frac{\mathrm{e}^{\sigma(\xi)}}{(1,\mathrm{e}^{\sigma})}, \quad \eta > \zeta,\end{aligned}\right\} \tag{8.45}$$

where $$(1,\mathrm{e}^{\lambda}) = \sum_x \exp\,\{\lambda(x)\}, \quad (1,\mathrm{e}^{\sigma}) = \sum_x \exp\,\{\sigma(x)\}$$

are inner products in the vector space of real functions on X. In particular

$$\sum_{\eta\leqslant\zeta}\hat{f}(\xi,\eta,\zeta) = f_A(\Gamma(*)\;\; \mathrm{e}^{\lambda(\xi)}/(1,\mathrm{e}^{\lambda}),$$

$$\sum_{\eta>\zeta}\hat{f}(\xi,\eta,\zeta) = f_A\{\Gamma(+)\,\mathrm{e}^{\lambda(\xi)}/(1,\mathrm{e}^{\sigma})$$

and $$\mathrm{e}^{\lambda(\xi)}/(1,\mathrm{e}^{\lambda}) = \hat{f}(\xi\,|\,\eta \leqslant \zeta), \quad \mathrm{e}^{\sigma(\xi)}/(1,\mathrm{e}^{\sigma}) = \hat{f}(\xi\,|\,\eta > \zeta)$$

are the surrogate conditional densities for profiles within deaths and survivors respectively. Finally the constants c_u^d and c_u^{s} in (8.44) are obtained by solving the two remaining designation equations, namely

$$(\psi_u^d, F_A) = F_A\{\Gamma(*)\}\sum_x \pi_u(x)\,\mathrm{e}^{\lambda(x)}/(1,\mathrm{e}^{\lambda}),$$

$$(\psi_u^{\mathrm{s}}, F_A) = F_A\{\Gamma(+)\}\sum_x \pi_u(x)\,\mathrm{e}^{\sigma(x)}/(1,\mathrm{e}^{\sigma}).$$

The surrogate bivariate density for profiles and lifetimes is

$$\hat{f}(\xi,\eta,\cdot) = f_A\{\Gamma(\eta,*)\}\frac{\mathrm{e}^{\lambda(\xi)}}{(1,\mathrm{e}^{\lambda})} + \sum_{\zeta<\eta}\frac{f_A\{\Gamma(+,\zeta)\}}{(T-\zeta)}\,\frac{\mathrm{e}^{\sigma(\xi)}}{(1,\mathrm{e}^{\sigma})} \tag{8.47}$$

and the surrogate marginal density for profiles alone is

$$\hat{f}(\xi,\cdot,\cdot) = f_A\{\Gamma(*)\}\frac{\mathrm{e}^{\lambda(\xi)}}{(1,\mathrm{e}^{\lambda})} + f_A\{\Gamma(+)\}\frac{\mathrm{e}^{\sigma(\xi)}}{(1,\mathrm{e}^{\sigma})}. \tag{8.47}$$

The ratio of these two quantities is $\hat{f}(\eta \mid \xi)$ the surrogate conditional density for lifetimes given the profile ξ and from it one can calculate a corresponding surrogate conditional survival function.

9. Foundations of data analysis

In the preceding sections the term 'data' has been used in the conventional sense of classical statistics. Now we examine the collection of data and its condensation at the deeper foundational level of Finch (1980*c*) to show that macrosurrogate spectra arise from computability, without appeal to equal informativeness, when data collection is non-disturbing in the precise sense defined below. While this adds little to the practical implementation of the data condensation procedures developed above, it does show that there is less room for developing alternative procedures than might appear at first sight. In a strictly logical approach the material of this section would be presented first and the preceding development modified accordingly.

9.1. *Pointers, readers and systems*

We examine the structure of data in terms of pointer-readings, namely the readings obtained by observational procedures which we call pointers. We start with two non-empty sets P and R whose elements are called unary pointers and unary readings respectively. An n-ary pointer is an ordered list $p = p_1 p_2 \dots p_n$ of n unary pointers. It denotes the consecutive use of its component unary pointers in the order $p_1, p_2, \dots, p_n$. Similarly an n-ary reading is an ordered list $r = r_1 r_2 \dots r_n$ of n unary readings. An n-ary pointer-reading is an ordered pair (p, r) consisting of an n-ary pointer $p = p_1 p_2 \dots p_n$ and an n-ary reading $r = r_1 r_2 \dots r_n$, where each component r_k is the unary reading on the corresponding unary pointer p_k. This interpretation of pointers and readings requires that their left-to-right ordering has a corresponding before-after temporal meaning and so, as in Finch (1980*c*), we should incorporate into the framework the times at which unary pointers are read. For our purposes here, however, it is sufficient to suppose that we are dealing with situations in which what we observe does not depend on when we observe it.

For brevity we refer to pointers and readings of arbitrary arity as multiple pointers and multiple readings. The set of all multiple readings is denoted by P_* and R_* denotes the set of all multiple readings. If p and q are multiple pointers, then we write $q < p$ and say that $p = p_1 p_2 \dots p_n$ extends q, or that q subtends p, when $q = p_1 p_2, \dots p_m$ for some $m < n$. Similarly we write $t < r$ when the multiple reading r extends the multiple reading t. If $p = p_1 p_2 \dots p_m$ and $q = q_1 q_2 \dots q_n$ are multiple pointers, then the multiple pointer $pq = p_1 p_2 \dots p_m q_1 q_2 \dots q_n$ is called the ordered juxtaposition of p and q. Similarly, rt denotes the ordered juxtaposition of the readings r and t.

From the viewpoint adopted here, the specification of a practical situation involves, *inter alia*, the enumeration of the multiple pointers which might then be used. The formal counterpart to this enumeration is called a system. This is a non-empty subset Ω of P_* such that all the subtending stages of a multiple pointer in Ω are themselves in Ω, in other words

$$p \in \Omega \;\&\; q < p \Rightarrow q \in \Omega. \tag{9.1}$$

It is an object that is defined by a set of possible observational procedures, any of which could be the one actually used on a given occasion. When we do observe the system Ω with a pointer p in Ω we obtain a corresponding reading r. The set of all pointer-readings (p, r), with p in Ω and r the corresponding reading when Ω is observed with p, is denoted by $PR(\Omega)$. Strictly speaking the r just mentioned names the reading which would be displayed by p because, in general, we observe a system with only one pointer. We suppose both that a pointer in Ω does not lead to more than one reading, that is

$$(p, r') \,\&\, (p, r'') \in PR(\Omega) \Rightarrow r' = r'', \tag{9.2}$$

and that the readings obtained from earlier stages of observation are not changed by later stages, that is

$$(q, t), (p, r) \in PR(\Omega) \,\&\, q < p \Rightarrow t < r. \tag{9.3}$$

It follows from (9.2) that the ordered pairs in $PR(\Omega)$ determine a function $s: \Omega \rightarrow R_*$. We write the functional symbol s to the right of its argument, so that ps denotes the unique reading for which (p, ps) is in $PR(\Omega)$. We say that s is the state of the system Ω. It follows from (9.3) that s preserves extension as well as arity. Any function from Ω to R_* which preserves arity and extension is a possible state of Ω.

9.2. *Observational disturbances and classical systems*

For p in the system Ω, let $\Omega_{|p}$ be the set of multiple pointers q such that pq is also in Ω. If $\Omega_{|p}$ is not empty, then it has the property (9.1) and is itself a system; it is called the conditional system determined by p. We say that p is terminal in Ω when $\Omega_{|p}$ is the empty set. For a non-terminal p and a state s of Ω, we define the function $s_{|p}: \Omega_{|p} \rightarrow R_*$, with domain $\Omega_{|p}$, by decreeing that, for each q in $\Omega_{|p}$, $qs_{|p}$ is the unique reading in R_* for which

$$(pq)\,s = (ps)(qs_{|p}). \tag{9.4}$$

It is easily verified that $s_{|p}$ preserves arity and extension, and it is therefore a state of $\Omega_{|p}$.

An ordered pair (Ω, s) conceptualizes the idea of a system Ω in a given state s. It is a deterministic concept because the system is defined by finite sequences of possible future observations and its state names the readings associated with each of them. When we observe ω in the state s with a non-terminal pointer p, the extensions of p in Ω determine a new system $\Omega_{|p}$ and a new state $s_{|p}$. Thus, in general, observation is accompanied by two disturbances: a change of system $\Omega \rightarrow \Omega_{|p}$ and a change of state $s \rightarrow s_{|p}$. If $\Omega_{|p} \neq \Omega$, then $s_{|p} \neq s$ because the two states are functions with different domains. But even if $\Omega_{|p} = \Omega$, then we will have $s_{|p} \neq s$ when there is q in Ω with $qs_{|p} \neq qs$. It is sometimes convenient to write sp for the state $s_{|p}$. Since $s_{|p}\,q = s_{|pq}$, we have

$$(sp)\,q = s(pq) \tag{9.5}$$

and, by an obvious extension of (9.4),

$$(p_1 p_2 \ldots p_n)\,s = (p_1 s)(p_2 \cdot sp_1)(p_3 \cdot sp_1 p_2) \ldots (p_n \cdot sp_1 p_2 \ldots p_{n-1}), \tag{9.6}$$

where $p_m \cdot sp_1 p_2 \ldots p_{m-1}$ is the unary reading from the unary pointer p_m when the state in question is $sp_1 p_2 \ldots p_{m-1}$.

A state s of the system Ω is said to be classical when $sp = s$ for each p in Ω. This means two things. Firstly, the states sp and s have the same domain, that is $\Omega_{|p} = \Omega$, and hence that $\Omega = Q_*$, the set of all finite strings generated by Q, the set of all the unary pointers belonging to Ω. Such systems are said to be simple systems. Only simple systems admit classical states. Secondly, $q(sp) = qs$ for all q and p in Ω, and so, from (9.6),

$$(p_1 p_2 \dots p_n)\, s = (p_1 s)(p_2 s) \dots (p_n s). \tag{9.7}$$

A simple system in a classical state is not disturbed by observation and, as shown by (9.7), the reading on a component unary pointer does not depend on what, if any, other unary pointers precede it.

Non-classical states occur explicitly in quantum mechanics (see, for example, Finch 1982*b*, 1984) and implicitly elsewhere, for instance in psychological experiments where a subject's response to a particular stimulus may depend on what other stimuli he has already experienced. But classical statistics has focused in the main on systems that can be considered, at least to a first approximation, as simple systems in classical states. This results in a number of important simplifications. Equation (9.7) shows that a classical state s of the simple system Q_* is determined by its associated unary state $sU = s \mid Q$, namely the restriction of the general state s to the set Q of its unary pointers. In particular, the system Q_* in the classical state s can be thought of as the ordered pair (Q, sU). In other words, the concept of a simple system in a classical state reduces to the familiar idea of a population of readings, namely the indexed set of unary readings $\{qs : q \in Q\}$, and hence to the idea of a data-set as defined at the beginning of the paper.

It is only a short step from the population concept to thinking of the unary readings qs as things which characterize the system Q_* in the state s, independently of whether or not we observe it. But that step fails to distinguish between what does not depend on how we observe it and what does not depend on whether we observe it. Strictly speaking, we are still involved with sequential observation when we are dealing with a simple system in a classical state because it is implicitly assumed that repetitions of some or all of the observations will lead to the same readings, at least to a first approximation.

Conversely a population of readings (Q, sU) may be thought of as a simple system in a classical state when each unary reading qs can be regarded as a fixed reading which is the reading on q at every occurrence of it, when we observe the system Q_* with a multiple pointer containing one or more occurrences of q as a component unary pointer.

9.3. *Indeterminism*

It was shown in Finch (1980*c*) that any system can be regarded as a simply system with an absorbing barrier. In what follows, therefore, we consider a simple system Q_* and for simplicity we suppose that its generating set of unary pointers $Q = \{q_1, q_2, \dots, q_N\}$ is a finite set. If Q_* is in the state s, then all the possible observationally induced states are determined, they are the sp with p in Q_*. Similarly, all the possible future readings are determined, they are the readings $q(sp)$ with both p and q in Q_*. In practice, however, the framework is indeterminate because certain states are indistinguishable at the observation level of practical enquiry. For while the actual state, a say, is a blueprint for what would be observed in all the observational futures in question, we are usually restricted to the use of a single observing pointer q in Q_*. Even though we observe the reading $r = qa$ we cannot, in general, thereby distinguish between the actual state a and other states s for which the corresponding reading qs is also r. In general, we can only name the actual state incompletely by saying that it is one of the states for which the reading given by observation with q is the one actually obtained. Thus indeterminism is the norm; it is its absence that is exceptional, not its presence. It is absent when the actual state is known to be classical and the observing pointer q is $q_1 q_2 \dots q_N$, because the observed reading is then $r_1 r_2 \dots r_N$ with $r_n = q_n a$. This determines the unary state aU and the full state a is then recoverable from the multiplicative formula (9.7). But, as we have seen earlier, even in this special case we might be restricted by choice, economy or force of circumstance to working with a summarizing condensation of the observed reading. Such condensations can also be used when the actual state is not classical.

Condensation for simple systems in possibly non-classical states can be formulated in a way which parallels our earlier discussion by noting that if the state is s and the observing pointer is q, then the associated reading qs is in R_* and may be condensed to $\delta(qs)$ by means of a condensing statistic δ, as defined in §1. The condensing statistic δ again determines a macrolevel, namely the equivalence $\rho = \delta^{-1}\delta$ on R_*, and this, together with the observing pointer q, determine the equivalence $I_{q,\rho}$ on the set of states under consideration which is given by the expression

$$s' I_{q,\rho}\, s'' \Leftrightarrow (qs')\,\rho(qs''). \tag{9.8}$$

States which are $I_{q,\rho}$-equivalent are indistinguishable at macrolevel ρ when the observing pointer is q.

While data condensation for systems in non-classical states is not our primary interest here, there are some general aspects of it which have bearing on the classical case. When the system Q_* is in a classical state s, interest focuses on the data spectrum, namely the multiplicities of the unary readings in the data-set

$$Qs = q_1 q_2 \dots q_N s = (q_1 s)(q_2 s) \dots (q_N s). \tag{9.9}$$

The data spectrum F_{Qs} condenses the unary state $sU\colon Q \rightarrow R$ by replacing it with the function F_{Qs} on R for which $F_{Qs}(r)$ is the number of pointers q in Q such that $qs = r$. But in the non-classical case the state of the system is no longer determined by its unary version and so, by itself, the data spectrum tells us little about the state of the system. The analogue of the data spectrum for the system Q_* in a non-classical state s is the condensation of the function $s\colon Q_* \rightarrow R_*$ given by the function K_s on R_* for which $K_s(r)$ is the number of pointers q in Q_* such that qs is r in R_*. We call K_s the spectrum of the state s. Its support B_s is the set of r in R_* at which $K_s(r) \neq 0$. If R_n is the set of n-ary readings in R_*, then

$$\sum_{R_n} K_s(r) = N^n. \tag{9.10}$$

The spectral density of the state s is the function k_s on R_* given by $k_s(r) = K_s(r)/N^{\alpha(r)}$, where $\alpha(r)$ is the arity of r. From (9.10)

$$\sum_{R_n} k_s(r) = 1, \quad n \geqslant 1. \tag{9.11}$$

When we are dealing with unary readings the spectrum and spectral density of the state s reduce to the data spectrum and data density, that is

$$\forall\, r \in R\colon K_s(r) = F_{Qs}(r), \quad k_s(r) = f_{Qs}(r). \tag{9.12}$$

If the state s is classical, then it follows from (9.7) that

$$\left.\begin{aligned} K_s(r_1 r_2 \dots r_n) &= K_s(r_1)\,K_s(r_2) \dots K_s(r_n), \\ k_s(r_1 r_2 \dots r_n) &= k_s(r_1)\,k_s(r_2) \dots k_s(r_n). \end{aligned}\right\} \tag{9.13}$$

These multiplicative rules for classical states should not be confused with the question of independence between variables. For instance, suppose the unary readings are themselves m-dimensional vectors $(x_1, x_2, \ldots, x_m)$. For each $j = 1, 2, \ldots, m$ we can construct the marginal density $f^{(j)}_{Qs}$ for the component variable x_j from the $f_{Qs}(x_1, x_2, \ldots, x_m)$ by summing over all the x_i with $i \neq j$. The component variables of the unary readings are independently distributed in the state s when

$$f_{Qs}(x_1, x_2, \ldots, x_m) = f^{(1)}_{Qs}(x_1) f^{(2)}_{Qs}(x_2) \ldots f^{(m)}_{Qs}(x_m).$$

This is quite a different requirement from (9.13) which comes simply from the fact that the state s there in question is a classical state.

For each positive integer n, the spectral density k_s of the state s is a density with finite support on the set of n-ary readings R_n. When s is a classical state, it follows from (9.13) and (4.46) that the likelihood of k_s at the multiple reading $r_1 r_2 \ldots r_n$ in the range of s is the value of k_s at that reading; that is

$$\text{lik}\,(k_s \,|\, r_1 r_2 \ldots r_n) = k_s(r_1 r_2 \ldots r_n). \tag{9.14}$$

Thus the phenomenological interpretation of the likelihood defined by (4.46) comes from the corresponding meaning of the spectral density of a classical state. For a non-classical we define the likelihood of k_s by equation (9.14).

Suppose that the actual state of the system Q_* is the possibly non-classical state a. When we observe Q_* with the multiple pointer q in Q_* and condense the reading on it at macrolevel ρ, the data condensation is $\delta(qa)$. The analogue of the procedure developed in the preceding sections is to talk about the system Q_* by means of a surrogate for the spectrum of its actual but unknown state which is computable from the condensation $\delta(qa)$. Surrogate state spectra are discussed in the next section. For simplicity we suppose that arity is part of the condensation.

9.4. *Surrogate state spectra*

A surrogate for the spectrum K_s of the state s of Q_* is a non-negative function $\tilde{K}$ on R_* with the same support as K_s and such that

$$\forall\, n \geqslant 1 : \sum_{R_n} \tilde{K}(r) = N^n. \tag{9.15}$$

An argument like that in §5.1 suggests that we should use a surrogate spectrum which has the same value at readings which are indistinguishable at the macrolevel of the condensation; in other words, that $\tilde{K}$ should be such that

$$\delta(r') = \delta(r'') \Rightarrow \tilde{K}(r') = \tilde{K}(r''). \tag{9.16}$$

For if $\delta(r') = \delta(r'')$ but $\tilde{K}(r') \neq \tilde{K}(r'')$, then the condensing statistic η given on R_* by $\eta(r) = (\delta(r), \tilde{K}(r))$ would distinguish between the ρ-equivalent readings r' and r'', and thereby call into question the correct identification of the macrolevel at which we were working. In what follows we suppose that a surrogate state spectrum based on the condensing statistic δ does have the property (9.16).

When the actual state is known to be classical, equation (9.13) suggests that its surrogate spectrum should have the multiplicative property

$$\tilde{K}(r_1 r_2 \ldots r_n) = \tilde{K}(r_1)\tilde{K}(r_2) \ldots \tilde{K}(r_n), \quad n \geqslant 1, \tag{9.17}$$

because the actual state spectrum for which $\tilde{K}$ is to deputize does have that property. We accordingly adopt (9.17) when we are dealing with classical states.

It follows from (9.16) and (9.17) that when the observing pointer is $q_1 q_2 \dots q_N$, a surrogate $\tilde{K}$ for a classical state spectrum K_s of the system Q_* is given on its unary support $S = B_s \cap R$, namely the support of the data-set Qs of (9.9), by the expression

$$\forall\, x \in S : \tilde{K}(x) = N\{Z(\psi)\}^{-1} \exp \psi(x), \quad \psi \in \Psi, \tag{9.18}$$

where

$$Z(\psi) = \sum_S \exp \psi(x), \tag{9.19}$$

and Ψ is the vector space of solutions to a system of linear equations determined by the condensing statistic. For the exponential form of (9.18) is trivially true with $\psi(x) = \ln\{C\tilde{K}(x)\}$ where the constant C is determined from equation

$$Z(\psi) = \sum_S \exp \psi(x) = C \sum_S \tilde{K}(x) = CN.$$

Equation (9.17) then shows that, for any N-ary data-set $D = r_1 r_2 \dots r_N$ with support S,

$$\tilde{K}(D) = N^N \{Z(\psi)\}^{-N} \exp\,[\sum_S F_D(x)\,\psi(x)]\,. \tag{9.20}$$

Finally, if ζ_ρ is the zeta function of the macrolevel $\rho = \delta^{-1}\delta$, that is $\zeta_\rho(D', D'')$ is 1 when $D'\rho D''$ and is 0 otherwise, then (9.16) gives

$$\zeta_\rho(D', D'') \sum_S \{F_{D'}(x) - F_{D''}(x)\}\,\psi(x) = 0 \tag{9.21}$$

and Ψ in (9.18) is the vector subspace of solutions to this system of linear equations.

It follows from (9.21) that

$$\Psi = \Phi(\rho),$$

where $\Phi(\rho)$ is the vector space of (5.13). Thus the unary restrictions of the surrogate spectra for classical states are the macrosurrogate spectra of (5.15). This results from (9.16) and (9.17) without invoking equal informativeness as presented in §5.2. If one adopts (9.16) and (9.17), then equal informativeness in the sense of (5.7) is a theorem, rather than an *ad hoc* principle, and one does not need to appeal to parsimony to justify the use of macrosurrogate spectra. The implication (9.16) is equivalent to requiring that on $\delta(K_s)$,

$$\tilde{K}(r) = H\{\delta(r)\}$$

is a function of the condensation of the multiple reading r and hence computable from it. This requirement is plausible because the underlying assumption is that multiple readings with the same condensation are being treated as if they were indistinguishable. Nevertheless it is difficult to argue that surrogate state spectra satisfying (9.16) are the only ones which are useful in practice. On the other hand, the use of a surrogate spectrum for a classical state which did not have the multiplicative property (9.17) would seem to conflict with the fact that the system in question was in a classical state. It follows from (9.17) that the associated surrogate state spectral density is also multiplicative, that is

$$\tilde{k}(r_1 r_2 \dots r_n) = \tilde{k}(r_1)\,\tilde{k}(r_2) \dots \tilde{k}(r_n),$$

and so it follows from (4.46) that, as at (9.14),

$$\text{lik}\;(\tilde{k}(|\, r_1 r_2 \dots r_n) = \tilde{k}(r_1 r_2 \dots r_n).$$

It follows from (9.21) that the surrogate spectra (9.18) are the same as those which would arise if the condensing statistic δ is the projection $\delta(D) = \mathbb{P}_{\Psi} F_D$. Since $\Psi = \Phi(\rho)$, this is the corollary to Theorem 5.4. If the condensing statistic δ is a prologjection, however, the results of this section suggest that one should use surrogate spectra which are macrosurrogate spectra for its projective closure $\bar{\delta}$ and not its metric-surrogate spectra. Thus the mathematical truth of the corollary to Theorem 7.4 does not mean that it is necessarily correct to use it in practice in the way suggested implicitly by the discussion in §7. In other words, as mentioned at the end of §7, the results of this section point to macrosurrogation rather than to metric-surrogation.

Finally we return to the point mentioned at the end of §4.6 *vis-à-vis* the formulation of the explanatory surrogation then under consideration as a generalized linear model and what we then called the natural link function. If we adopt (9.16) and (9.17), then we may use only the surrogate spectra in $\mathscr{E}(\Phi)$ and, correspondingly, we must then use the natural link function. The use of other link functions to construct surrogate spectra would violate (9.16). This does not mean, of course, that such surrogate spectra cannot be useful in practice.

9.5. *Concluding remarks*

If a pointer is chosen at random from the set of n-ary pointers in Q_* when that system is in the state s, then $k_s(r_1 r_2 \ldots r_n)$ is the probability that the reading on that pointer is $r_1 r_2 \ldots r_n$. In such a sampling context it is little more than a matter of taste whether one calls k_s a state spectral density or a probability density. But the probabilistic terminology, by itself, adds little to our understanding of the elusive concept of probability. It is simply a special case of the fact that a concept of probability already defined on pointers can be transferred to the readings on them. Moreover a system, as defined above, does not involve probabilities on its pointers. Introducing probabilities on pointers would lead to the concept of a probabilistic system. Whilst such a concept might be useful in some contexts, it is not the one under study here. In the data framework of this section, the term 'system' conceptualizes the idea of a physical object as defined by the totality of ways in which we might observe it, without regard to the relative frequencies with which we might adopt the various observing procedures then in question. If probabilities were introduced into the data framework to mimic real-world indeterminacy, then they would enter as probabilities on states, not pointers.

Since a state spectrum is a list of counts it endows data analysis with a structure which is similar to that of the probability calculus. The similarity lies not only in the use of counting measures *per se*, but also in the practical motivation for their use. For example, at the beginning of §6 we noted that the assessment of surrogate performance involves not only how well a proposed surrogate depicts the suppressed spectrum, the issue discussed in that section, but also how effective similar condensation might be for other data-sets. Consider the case in which the state s is classical, the observing pointer is $q_1 q_2 \ldots q_N$, so that the actual data-set $A = Qs$ of (9.9), and the other data-sets in mind are those which would arise had we observed Q_* in the same state s with a different pointer, say the n-ary pointer q in Q_* where $1 \leqslant n \leqslant N$. Suppose that $\hat{F}_{Qs|\rho}$ is a good depiction of F_{Qs}, that is the information deviance $\Delta[F_{Qs}, \hat{F}_{Qs|\rho}]$ is relatively small and the percentage reduction in deviance is correspondingly high. One could examine how effective depiction at level ρ is for the other data sets in mind by working out how $\Delta[F_D, \hat{F}_{D|\rho}]$ varies with $D = qs$ and q

running through Q_n, the set of all n-ary pointers in Q_*. The deviance $\Delta[F_D, \hat{F}_{D|\rho}]$ is a function of the reading $D = qs$ and the relative frequencies of those readings are given by $k_s \,|\, Q_n$, the restriction of the spectral density of s to Q_n. Thus, because of (9.13), the relative frequency distribution of the deviance can be regarded as its sampling density over independent, with replacement, ordered random samples from a population of size N whose density is the suppressed data density f_A. In other words the question of how effective similar condensation might be for other data-sets can be investigated by a bootstrap calculation. If we replace k_s in that calculation by its surrogate $\hat{k}_s$, and regard it as an estimate of an underlying population-based likelihood, then we obtain the corresponding estimate of the sampling density of the deviance.

Bootstrap calculation arises here, not as a substitute for a corresponding probability calculation, but as a meaningful way of investigating, within a purely data analytic framework, questions about the general usefulness of data condensation procedures. This suggests, but does not prove, that probability need not play as great a role in applied statistics as is sometimes supposed.

Nevertheless it is perhaps debatable whether such a bootstrap calculation addresses the point at issue in the only meaningful way. It can be argued, for example, that it is the magnitude of the change in the deviance $\Delta[F_D, \hat{F}_{D|\rho}]$ when D is perturbed that is of primary interest, and that the multiplicity of the pointers q in Q_* with $qs = D$ is irrelevant. From that point of view it is also informative to examine how the deviance changes as D runs uniformly through the n-ary readings in the range of s, namely the support of $k_s \,|\, Q_n$.

References

Barndorff-Nielson, O. 1978 *Information and exponential families*. New York: Wiley.

Bliss, C. I. 1935 The calculation of the dosage–mortality curve. *Ann. appl. Biol.* **22**, 134–167.

Campbell, L. L. 1970 Equivalence of Gauss's principle and minimum discrimination of probabilities, *Ann. math. Statist.* **41**, 1011–1013.

Cox, D. R. 1972 Regression models and life-tables. *Jl R. statist. Soc.* B **34**, 187–219.

Dobson, A. J. 1983 *An introduction to statistical modelling*. London: Chapman and Hall.

Dutta, M. 1966 On maximum (information-theoretic) entropy estimation. *Sankhya* A **28**, 319–321.

Finch, P. D. 1977 Rough analysis and spurious accuracy. *Aust. J. Statist.* **19**, 1–21.

Finch, P. D. 1980*a* The probability grating problem. *J. management Sci. appl. Cyb.* **8**, 16–20.

Finch, P. D. 1980*b* Flat approximation to the distribution of extreme values and its use in the statistical treatment of floods. In *Applied cybernetics* (ed. A. Ghosal). New Delhi: South Asian Publishers.

Finch, P. D. 1980*c* The formal structure of observational procedures. In *Semigroups* (ed. P. Hall, P. Jones & G. Preston). Academic Press.

Finch, P. D. 1981 On the role of description in statistical enquiry. *Br. J. Phil. Sci.* **32**, 127–144.

Finch, P. D. 1982*a* Invariate surrogation in scientific dialogue. *Acta. Math. Scientia* **2**, 149–178.

Finch, P. D. 1982*b* Classical probability and the quantum mechanical trace formulation of quantum mechanics. *Found. Phys.* **12**, 327–345.

Finch, P. D. 1984 The operator formalism of quantum mechanics from the viewpoint of short disturbances in non-relativistic classical motion. *Found. Phys.* **14**, 281–306.

Kullback, S. & Liebler, R. A. 1951 On information and sufficiency. *Ann. math. Statist.* **22**, 79–86.

Rockafellar, R. P. 1970 *Convex analysis*. Princeton University Press.

Received 22 March 1990; accepted 3 October 1990