Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems

Imre Csiszar

The Annals of Statistics, Vol. 19, No. 4. (Dec., 1991), pp. 2032-2066.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199112%2919%3A4%3C2032%3AWLSAME%3E2.0.CO%3B2-J

The Annals of Statistics is currently published by Institute of Mathematical Statistics.

The Annals of Statistics
1991, Vol. 19, No. 4, 2032–2066

WHY LEAST SQUARES AND MAXIMUM ENTROPY? AN AXIOMATIC APPROACH TO INFERENCE FOR LINEAR INVERSE PROBLEMS[1]

BY IMRE CSISZÁR

Mathematical Institute of the Hungarian Academy of Sciences

An attempt is made to determine the logically consistent rules for selecting a vector from any feasible set defined by linear constraints, when either all n-vectors or those with positive components or the probability vectors are permissible. Some basic postulates are satisfied if and only if the selection rule is to minimize a certain function which, if a "prior guess" is available, is a measure of distance from the prior guess. Two further natural postulates restrict the permissible distances to the author's f-divergences and Bregman's divergences, respectively. As corollaries, axiomatic characterizations of the methods of least squares and minimum discrimination information are arrived at. Alternatively, the latter are also characterized by a postulate of composition consistency. As a special case, a derivation of the method of maximum entropy from a small set of natural axioms is obtained.

1. Introduction. A frequently occurring problem in statistics and applied mathematics is that a function has to be inferred from insufficient information that specifies only a feasible set of functions. Problems of this kind are often called inverse problems. Typical examples are the reconstruction of a signal or image from the results of certain measurements, and the assignment of a probability density or mass function subject to moment constraints, that is, constraints that specify the expectations of certain functions of the underlying random variable. Often the practical solution to such problems is to select an element of the feasible set by a more or less ad hoc rule, usually by minimizing some functional such as the L_2-norm or negative entropy. If some function is specified as a "prior guess," it is natural to minimize a measure of distance from the latter, most often the L_2-distance, or, for probability density or mass functions, Kullback's I-divergence (also called information for discrimination or cross-entropy).

In this paper we address the question of what selection rules are "good" in this framework and, in particular, whether the mentioned standard ones are indeed the "best." Unfortunately, it is hard to give a mathematical meaning to this question. It does not seem possible to define a general criterion by which

Received May 1989; revised December 1990.

[1]This research was started during the author's visit at the University of Tokyo, supported by the Japan Society for the Promotion of Science. Its completion was supported by the Hungarian National Science Foundation, Scientific Research Grant 1806.

AMS 1980 *subject classifications.* Primary 62A99; secondary 68T01, 94A17, 92C55.

Key words and phrases. Image reconstruction, linear constraints, logically consistent inference, minimum discrimination information, nonlinear projection, nonsymmetric distance, selection rule.

the goodness of selection rules could be compared. Still, whereas people apparently feel little need for any special justification of least squares (L_2-norm minimization), various reasons have been put forward to justify I-divergence minimization [introduced into statistics by Kullback (1959) as the method of minimum discrimination information] and entropy maximization. The recent widespread applications of "maximum entropy" have been pioneered to a great extent by Jaynes [for his views cf. Jaynes (1982)]. The present author has argued elsewhere [Csiszár (1985)] that the conditional limit theorems of Van Campenhout and Cover (1981) and Csiszár (1984) suggest the interpretation that the minimum I-divergence "updating" of a prior probability distribution to meet moment constraints is a limiting form of Bayesian updating.

Here we adopt an axiomatic approach and consider those selection rules as "good" that lead to a logically consistent method of inference, in the sense of satisfying some natural postulates. The term inference is not meant in a statistical sense. Indeed, our considerations will be nonprobabilistic, even though the objects to be inferred may be probability distributions. The relation of this work to previous axiomatic approaches will be discussed at the end of this section.

As a typical example, we briefly sketch a model of the image reconstruction problem that occurs in computerized tomography and in various other fields [cf., e.g., Herman and Lent (1976) or Censor (1983)]. An image is represented by a positive-valued function f defined on some domain. The available information consists in the measured values of some linear functionals $R_i f$, $i = 1, \ldots, k$. In X-ray tomography, f is the unknown X-ray attenuation function and $R_i f$ is its integral along the path of the i's ray. Now, the domain of f is partitioned into a finite number of picture elements, called pixels, numbered in some way from 1 to n. Assuming that f is nearly constant within each pixel, we can write

$$f = \sum_{j=1}^{n} v_j f_j, \tag{1.1}$$

where f_j is the indicator function of the jth pixel. Then, setting $a_{ij} = R_i f_j$ and $b_i = R_i f$, we have

$$\sum_{j=1}^{n} a_{ij} v_j = b_i, \quad i = 1, \ldots, k. \tag{1.2}$$

Thus the unknown function f is represented by the vector $\mathbf{v} = (v_1, \ldots, v_n)^T$, and the feasible set is identified with the set of those vectors $\mathbf{v}$ with positive components that satisfy the linear constraints (1.2). The reconstruction problem is to select a "suitable" element of this feasible set (possibly depending on a "prior guess" of f represented by a vector $\mathbf{u}$).

In this paper, we concentrate on linear inverse problems of form (1.2). The extension of our results to the continuous case, that is, to inferring functions defined on some domain and not representable by finite-dimensional vectors, should not be hard but will not be entered. The above example will be

repeatedly used as an illustration, but it should be emphasized that our axiomatic approach is by no means limited to image reconstruction. On the other hand, since a general approach inevitably involves idealizations, the solutions it leads to are not necessarily "best" for specific practical problems, including image reconstruction.

Our goal is to determine the "logically consistent" rules for selecting an element from any possible feasible set. We adopt the idealized assumption that all conceivable linear constraints may occur; thus the possible feasible sets are all those subsets of a basic set S of permissible vectors that can be defined by constraints of form (1.2). Three cases will be considered in a parallel manner namely when S consists of all n-vectors, or of those with positive components (as in image reconstruction) or of the probability vectors with positive components. The choice of vectors with positive rather than nonnegative components has been preferred in order to reduce technical difficulties; also, this ensures that a nonnegative quantity is never inferred to be 0 when the available information permits it to be positive, which is generally considered desirable.

Our postulates will be stated and intuitively justified in Section 2. The results will be stated in Section 3 and proved in Section 5, using the lemmas in Section 4. The key result is Theorem 1, namely that the basic postulates of "regularity" and "locality" of a selection rule are satisfied if and only if the selection is by minimizing a certain function. If a prior guess is available, this function is a measure of distance (nonsymmetric in general) from the prior guess. The subsequent theorems show how certain additional postulates restrict the class of functions that may be used. Perhaps the most striking result in Theorem 5. It provides a parallel characterization of the methods of least squares and minimum I-divergence as the only ones satisfying, in addition to regularity and locality, a postulate of "composition consistency." The intuitive meaning of this postulate is that if the object of inference is composed of two components, and the available information says nothing about their interaction, we should infer that no interaction is present.

It should be mentioned that in practice (1.2) is often relaxed to

$$\sum_{j=1}^{n} a_{ij} v_i + e_i = b_i, \qquad i = 1, \ldots, k, \tag{1.3}$$

where $\mathbf{e} = (e_1, \ldots, e_k)^T$ is an error vector. For the reconstruction problem of positron emission tomography, Vardi, Shepp and Kaufman (1985) described a model equivalent to (1.3) with data b_i that were Poisson random variables (counts of detected emissions in "tubes" determined by pairs of detectors). A vector $\mathbf{e}$ as in (1.3) did not explicitly enter their model, but, defining e_i as the difference of the actual and expected counts for the ith tube, (1.3) would hold. The suggested reconstruction was the maximum likelihood estimate of $\mathbf{v}$, and for its computation the EM algorithm of Dempster, Laird and Rubin (1977) was used. This reconstruction technique is sometimes considered as related to "maximum entropy" [cf., e.g., Miller and Snyder (1987)], for the formal rather than conceptual reason that the EM algorithm is equivalent to an alternating

I-divergence minimization; Vardi, Shepp and Kaufman (1985) have shown that the convergence of their algorithm is an instance of a result of Csiszár and Tusnády (1984) on alternating I-divergence minimization.

Maximum likelihood estimation is not a generally applicable method for "solving" inverse problems of form (1.3) because (i) the "errors" e_i may be nonrandom or else their joint distribution may be unknown and (ii) even if the "errors" are random with known distribution, the maximum likelihood estimate is typically nonunique if $n > k$. Vardi, Shepp and Kaufman (1985) avoided the latter difficulty by partitioning the object into fewer pixels than the number of "tubes" that was sufficiently large. If bounds on the magnitude of the errors are known, it may be convenient to interpret (1.3) as a system of inequalities of form

$$b_i' \le \sum_{j=1}^{n} a_{ij} v_j \le b_i'', \qquad i = 1, \ldots, k, \tag{1.4}$$

or as an inequality of form

$$\sum_{i=1}^{k} \left(\sum_{j=1}^{n} a_{ij} v_j - b_i \right)^2 \le c. \tag{1.5}$$

Our axiomatic approach could be extended to the problem of selecting an element from any set defined by inequality constraints such as (1.4) or (1.5) or, more generally, from any convex subset of the basic set S. Alternatively, (1.3) could be interpreted as determining a "feasible set" of pairs $(\mathbf{v}, \mathbf{e})$, and a selection from this set could be made by any method that has been deemed "good" for the problem (1.2). Indeed, this is often done in practice, for example, by minimizing some quadratic function of the pair $(\mathbf{v}, \mathbf{e})$ [cf. Herman and Lent (1976) or Censor (1983)]. It remains to be seen whether "solutions" of this kind to the problem (1.3) can be covered by an extension of our axiomatic approach; one of the difficulties is that the possible linear constraints on pairs $(\mathbf{v}, \mathbf{e})$ are of the very special form (1.3).

The approach in this paper was strongly motivated by Shore and Johnson (1980), where—for the problem of assigning a probability distribution subject to moment constraints—an intuitively appealing axiomatic derivation of the methods of maximum entropy and minimum I-divergence was provided. Skilling (1988) gave a similar derivation for inferring arbitrary positive-valued functions; because of the greater liberty, this case turned out to be considerably simpler. Both Shore and Johnson (1980) and Skilling (1988) started from the assumption that inference was based on minimizing some function or, equivalently, on some transitive ranking that could be described by real numbers. After having submitted this paper, the author learned that Paris and Vencovská (1990) had arrived at "the inevitability of maximum entropy" from axioms that—like ours—did not a priori assume the minimization of some function; in other respects, our approach appears quite different from theirs, although similarities in the axioms do exist. The author is indebted to Professor Paris for sending him manuscripts of this and related works.

Our results also provide axiomatic characterizations of measures of distance whose minimization leads to "good" methods of inference. Whereas there is an extensive literature on axiomatic characterizations of entropy, I-divergence and their generalizations [cf., e.g., Aczél and Daróczy (1975)], our characterizations substantially differ from the usual ones: Our postulates involve not the measure to be characterized but rather the inference method it leads to. In Theorem 2(ii), a class of distances introduced by Csiszár (1963) [and independently by Ali and Silvey (1966)] is characterized; related results also appear in Shore and Johnson (1980) and Skilling (1988). Theorem 3 characterizes a class of distances introduced by Bregman (1967); since this paper had been submitted, a different axiomatic characterization of this class (in the continuous case) was given by Jones and Byrne (1990). Finally, we comment on the one-parameter family of distances characterized in Theorem 4(ii). In the original version of this paper, the previously not considered members of that family had been mentioned as candidates for becoming practically useful. Recently, Jones and Trutzer (1989) reported applications of (continuous versions of) these distances that seem to confirm that prediction.

2. Preliminaries, postulates. The real line and the positive half-line are denoted by R and R_+, respectively. We emphasize that R_+ does not contain 0. The vectors in R^n whose components are all 0 or all 1 are denoted by $\mathbf{0}$ or $\mathbf{1}$. All vectors are column vectors.

The set of n-dimensional vectors with positive components of sum 1 is denoted by Δ_n, that is,

$$\Delta_n = \{\mathbf{v}: \mathbf{v} \in R_+^n, \mathbf{1}^T\mathbf{v} = 1\}. \tag{2.1}$$

The family of affine subspaces of R^n, that is, the family of nonvoid subsets of R^n defined by linear constraints, will be denoted by $\mathscr{L}_n$. In other words $L \in \mathscr{L}_n$ iff $L \neq \emptyset$ and

$$L = \{\mathbf{v}: \mathbf{v} \in R^n, \mathbf{A}\mathbf{v} = \mathbf{b}\} \tag{2.2}$$

for some $\mathbf{A} \in M_{k\times n}$ ($k \times n$ matrix) and $\mathbf{b} \in R^k$. We denote by $\mathscr{L}_n^+$ the family of all nonvoid subsets of R_+^n of form (2.2), with $\mathbf{v} \in R^n$ replaced by $\mathbf{v} \in R_+^n$. The family of those $L \in \mathscr{L}_n^+$ which are subsets of Δ_n will be denoted by $\mathscr{L}_n^+(1)$. For any fixed dimension $d < n$, the set of d-dimensional elements of $\mathscr{L}_n$ (or $\mathscr{L}_n^+$) is endowed with the natural topology, that is, the quotient topology derived from the Euclidean topology of the set of those pairs $(\mathbf{A}, \mathbf{b}) \in M_{(n-d)\times n} \times R^{n-d}$ that define a d-dimensional element of $\mathscr{L}_n$ (or $\mathscr{L}_n^+$).

Throughout this paper, our basic set S will be either of R^n, R_+^n and Δ_n. The set of components of vectors in S will be denoted by V, that is, V stands for R, R_+ or the open interval $(0, 1)$, according to the choice of S. Unless stated otherwise, u, v and w will always denote elements of V, whereas $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ denote vectors in V^n. Further, we denote by $\mathscr{L}$ the family of nonvoid subsets of S determined by linear constraints. Thus, according to the three cases, $\mathscr{L}$ equals $\mathscr{L}_n$, $\mathscr{L}_n^+$ or $\mathscr{L}_n^+(1)$. For convenience, we will call the elements of $\mathscr{L}$ subspaces of S also if $S \neq R^n$.

Notice that S itself is an element of $\mathscr{L}$; among the proper subsets of S the maximal subspaces are those of dimension $n-1$ (if $S = R^n$ or R_+^n) or $n-2$ (if $S = \Delta_n$). The subfamily of $\mathscr{L}$ consisting of these maximal subspaces will be denoted by $\mathscr{M}$. Thus $\mathscr{M}$ consists of the (nonvoid) sets

$$L = \begin{cases} \{\mathbf{v}: \mathbf{a}^T\mathbf{v} = b\},\ \mathbf{a} \neq 0, & \text{if } S = R^n \text{ or } R_+^n, \\ \{\mathbf{v}: \mathbf{a}^T\mathbf{v} = b,\ \mathbf{1}^T\mathbf{v} = 1\},\ \mathbf{a} \neq \lambda\mathbf{1}, & \text{if } S = \Delta_n. \end{cases} \tag{2.3}$$

As mentioned previously, the condition $\mathbf{v} \in V^n$ is implicit in the notation in (2.3).

Having in mind inference problems as in Section 1, we are interested in rules that specify for each $L \in \mathscr{L}$ (and "prior guess" $\mathbf{u}$) an element of L, to be selected if L is in the feasible set (and the prior guess was $\mathbf{u}$). The vector selected from L when a prior guess $\mathbf{u}$ was available is regarded as a nonlinear projection on L of $\mathbf{u}$, denoted by $\Pi(L|\mathbf{u})$.

DEFINITION 1. A *selection rule* (with basic set S) is a mapping $\Pi: \mathscr{L} \to S$ such that $\Pi(L) \in L$ for every $L \in \mathscr{L}$. A *projection rule* is a family of selection rules $\Pi(\cdot|\mathbf{u})$, $\mathbf{u} \in S$, such that $\mathbf{u} \in L$ implies $\Pi(L|\mathbf{u}) = \mathbf{u}$. A selection rule is *generated by a function* $F(\mathbf{v})$, $\mathbf{v} \in S$, if for every $L \in \mathscr{L}$, $\Pi(L)$ is the unique element of L where $F(\mathbf{v})$ is minimized subject to $\mathbf{v} \in L$. A projection rule is *generated by a function* $F(\mathbf{v}|\mathbf{u})$, $\mathbf{u} \in S$, $\mathbf{v} \in S$, if its component selection rules $\Pi(\cdot|\mathbf{u})$ are generated by the functions $F(\cdot|\mathbf{u})$.

REMARK. A necessary condition for $F(\mathbf{v}|\mathbf{u})$ to generate a projection rule is that for any fixed $\mathbf{u} \in S$ the unique global minimum of F on S be attained at $\mathbf{v} = \mathbf{u}$. Attention may be restricted to functions F for which this minimum is 0 [because a projection rule generated by some $\tilde{F}(\mathbf{v}|\mathbf{u})$ is also generated by $F(\mathbf{v}|\mathbf{u}) = \tilde{F}(\mathbf{v}|\mathbf{u}) - \tilde{F}(\mathbf{u}|\mathbf{u})$]. A function $F(\mathbf{v}|\mathbf{u})$, $\mathbf{u} \in S$, $\mathbf{v} \in S$, with the property $F(\mathbf{v}|\mathbf{u}) \geq 0$, with equality iff $\mathbf{v} = \mathbf{u}$, will be called a *measure of distance* on S.

EXAMPLE 1. In the case $S = R^n$, the Euclidean distance $F(\mathbf{v}|\mathbf{u}) = \|\mathbf{u} - \mathbf{v}\|$ generates a projection rule that gives rise to the ordinary projection in Euclidean geometry. It will be called the *least squares projection rule* and its component selection rules will be called *least squares selection rules*. In particular, the selection rule generated by $F(\mathbf{v}) = \|\mathbf{v}\|$ is the *standard least squares selection rule*. A projection rule generated by a weighted L_2-distance $(\sum_{i=1}^n a_i(v_i - u_i)^2)^{1/2}$ will be called a *weighted least squares projection rule*. Of course, "least squares" is a standard method of a very long history. Notice, however, that it is not suitable in the cases $S = R_+^n$ or Δ_n; then $F(\mathbf{v}) = \|\mathbf{v} - \mathbf{u}\|$ does not generate a selection rule for any $\mathbf{u} \in S$ (neither does any weighted L_2-distance) because it does not attain a minimum on some subspaces $L \in \mathscr{L}$.

EXAMPLE 2. Let $S = R_+^n$ or Δ_n. The I-divergence of $\mathbf{v} \in S$ from $\mathbf{u} \in S$ is defined by

$$I(\mathbf{v}\|\mathbf{u}) = \sum_{i=1}^n \left(v_i \log \frac{v_i}{u_i} - v_i + u_i \right). \tag{2.4}$$

This generalization of Kullback's formula

$$I(\mathbf{v}\|\mathbf{u}) = \sum_{i=1}^{n} v_i \log \frac{v_i}{u_i}, \qquad \mathbf{u} \in \Delta_n, \mathbf{v} \in \Delta_n,$$

retains the property $I(\mathbf{v}\|\mathbf{u}) \geq 0$, with equality if and only if $\mathbf{u} = \mathbf{v}$. Clearly, $F(\mathbf{v}|\mathbf{u}) = I(\mathbf{v}\|\mathbf{u})$ generates a projection rule (for any fixed $\mathbf{u} \in S$, it attains a unique minimum on any $L \in \mathscr{L}$, and if $\mathbf{u} \in L$, this minimum is attained at $\mathbf{v} = \mathbf{u}$); this will be called the *I-divergence projection rule*, and either of its component selection rules is an *I-divergence selection rule*. The selection rule generated by the negative entropy $F(\mathbf{v}) = \sum_{i=1}^{n} v_i \log v_i$ will be called the *maximum entropy selection rule*. This is a special case of I-divergence selection rules corresponding to $\mathbf{u} = (1/n)\mathbf{1}$ if $S = \Delta_n$ and $\mathbf{u} = (1/e)\mathbf{1}$ if $S = R_+^n$. As a comparison of the naturality of these choices of $\mathbf{u}$ indicates, the maximum entropy selection rule plays a distinguished role mainly in the case $S = \Delta_n$, that is, for inferring probability distributions.

We emphasize that we do not a priori restrict attention to selection rules generated by some function. However, the postulates in Definitions 2 and 3 below will suffice to prove such generatedness.

Given a selection rule Π with basic set S, we will designate $\Pi(S)$ by $\mathbf{v}^0(\Pi)$. Then the component selection rules $\Pi(\cdot|\mathbf{u})$ of a projection rule satisfy, by Definition 1,

$$\mathbf{v}^0(\Pi(\cdot|\mathbf{u})) = \mathbf{u}. \tag{2.5}$$

DEFINITION 2. A selection rule $\Pi: \mathscr{L} \to S$ is *regular* if it satisfies the following axioms:

(i) (consistency) if $L' \subset L$ and $\Pi(L) \in L'$, then $\Pi(L') = \Pi(L)$;

(ii) (distinctness) if $L \neq L'$ are both in $\mathscr{M}$, then $\Pi(L) \neq \Pi(L')$ unless both L and L' contain $\mathbf{v}^0(\Pi)$;

(iii) (continuity) the restriction of Π to the subspaces of any fixed dimension is continuous.

A projection rule is *regular* if its component selection rules are such.

The consistency axiom formalizes the intuitive idea that if $\mathbf{v} = \Pi(L)$ selected on the basis of constraints specifying L also satisfies the stronger constraints specifying L', then the additional constraints provide no reason to change the selection of $\mathbf{v}$. The case for this axiom appears quite strong, for example, in image reconstruction one would hardly want to use a selection rule not satisfying it. Nevertheless, this axiom may be inappropriate for certain problems. When the selected element ought to resemble the other elements of the feasible set L as much as possible, it is reasonable to select the $\mathbf{v} \in L$ that minimizes—for some given measure of distance d—either $\sup_{\mathbf{w} \in L} d(\mathbf{v}, \mathbf{w})$ ["barycenter method," Perez (1984)] or the conditional expectation of $d(\mathbf{v}, \mathbf{w})$ on the condition $\mathbf{w} \in L$ (Bayesian rule, requires a prior distribution on S).

These selection rules do not satisfy the consistency axiom and are outside the scope of this paper.

The distinctness axiom says that different information, both consisting in a single linear constraint, must lead to different conclusions, unless both are consistent with the selection that would be made without any constraints. This is a technical postulate and it would be desirable if it could be dispensed with. Notice that the distinctness axiom is certainly satisfied for selection rules generated by differentiable functions F.

Continuity is an obvious regularity hypothesis. Notice, however, that (for projection rules) we did not postulate the continuous dependence of $\Pi(L|\mathbf{u})$ on $\mathbf{u}$.

For any set of indices $J = \{j_1, \dots, j_k\}$, $1 \le j_1 < \cdots < j_k \le n$, and any vector $\mathbf{a} \in R^n$, we denote by $\mathbf{a}_J$ the vector in R^k defined by

$$\mathbf{a}_J = (a_{j_1}, \dots, a_{j_k})^T. \tag{2.6}$$

For a selection rule Π, we will denote $(\Pi(L))_J$ briefly by $\Pi_J(L)$.

DEFINITION 3. A selection rule $\Pi\colon \mathscr{L} \to S$ is *local* if for every $J \subset \{1, \dots, n\}$ of arbitrary size k, and any L' and L'' in $\mathscr{L}$ of form

$$L' = \{\mathbf{v}\colon \mathbf{v}_J \in L_0, \mathbf{v}_{J^c} \in \tilde{L}'\}, \qquad L'' = \{\mathbf{v}\colon \mathbf{v}_J \in L_0, \mathbf{v}_{J^c} \in \tilde{L}''\}, \tag{2.7}$$

where $L_0 \in \mathscr{L}_k$, $\tilde{L}' \in \mathscr{L}_{n-k}$, $\tilde{L}'' \in \mathscr{L}_{n-k}$ (if $S = R^n$) or $L_0 \in \mathscr{L}_k^+$, $\tilde{L}' \in \mathscr{L}_{n-k}^+$, $\tilde{L}'' \in \mathscr{L}_{n-k}^+$ (if $S = R_+^n$ or Δ_n), we have

$$\Pi_J(L') = \Pi_J(L''). \tag{2.8}$$

A projection rule is *local* if its component selection rules $\Pi(\cdot|\mathbf{u})$ are local and, in addition, for L' and L'' as above we have

$$\Pi_J(L'|\mathbf{u}') = \Pi_J(L''|\mathbf{u}'') \quad \text{if } \mathbf{u}'_J = \mathbf{u}''_J. \tag{2.9}$$

REMARK. If $S = R^n$, $\mathscr{L} = \mathscr{L}_n$ or $S = R_+^n$, $\mathscr{L} = \mathscr{L}_n^+$, then any L' and L'' of form (2.7) necessarily belong to $\mathscr{L}$. On the other hand, if $S = \Delta_n$, $\mathscr{L} = \mathscr{L}_n^+(1)$, the sets L' and L'' in (2.7) belong to $\mathscr{L}$ iff the sum of components of each vector in L_0 is equal to the same $0 < c < 1$, and the sum of components of each vector in $\tilde{L}'$ and $\tilde{L}''$ is equal to $1 - c$.

Locality means, in other words, that for $L' \in \mathscr{L}$ as in (2.7), $\Pi_J(L')$ depends only on L_0, and $\Pi_J(L'|\mathbf{u}')$ depends only on L_0 and $\mathbf{u}'_J$.

Intuitively, the locality axiom says that whenever the available information consists of two pieces that involve complementary sets of components of the vector to be inferred, each component of the vector selected on the basis of this information will be determined by that piece of information which involves the component in question. In the reconstruction problem of X-ray tomography, this means that if two sets of ray paths covering disjoint sets of pixels were used, then for each pixel the reconstruction would be determined by those rays whose paths are in the set covering that pixel. This axiom appears very

natural, but strict adherence to it may perhaps be criticized in the tomography example on the basis of smoothness properties of the unknown X-ray attenuation function. For the case of inferring probability mass functions, Shore and Johnson (1980) used a similar but stronger postulate called "subset independence."

In this paper, regularity and locality will be the basic postulates. The selection and projection rules satisfying them will be characterized in Theorem 1. In the rest of this section we formulate some other desirable properties that suggest themselves as additional postulates, if we want to arrive at methods of practical interest.

An important role will be played by the special subspaces

$$L_{ij}(t) = \begin{cases} \{\mathbf{v}: v_i + v_j = t\}, & \text{if } S = R^n \text{ or } R_+^n, \\ \left\{\mathbf{v}: v_i + v_j = t, \sum_{l \neq i,j} v_l = 1 - t\right\}, & \text{if } S = \Delta_n. \end{cases} \tag{2.10}$$

Given a selection rule Π, we will write $v \leftrightarrow_{ij} v'$ to designate that v and v' equal the ith and jth components of $\Pi(L_{ij}(t))$ for some t; of course, then necessarily $t = v + v'$. Similarly, given a local projection rule, we will write $(v|u) \leftrightarrow_{ij} (v'|u')$ to designate that v and v' equal the ith and jth components of $\Pi(L_{ij}(t)|\mathbf{u})$ if the ith and jth component of $\mathbf{u}$ are u and u' (and $t = v + v'$).

Clearly, $v \leftrightarrow_{ij} v'$ means the same as $v' \leftrightarrow_{ji} v$. Further, we will show (corollary of Lemma 3) that if $v \leftrightarrow_{ij} v'$ and $v' \leftrightarrow_{jk} v''$ for a regular, local selection rule Π, then also $v \leftrightarrow_{ik} v''$, provided in the case $S = \Delta_n$ that $v + v' + v'' < 1$. Of course, similar statements hold for the relations $(v|u) \leftrightarrow_{ij} (v'|u')$ associated with a regular, local projection rule.

DEFINITION 4. (i) A local projection rule is *semisymmetric* if for every $L_{ij}(t)$ as in (2.10), the ith and jth components of $\Pi(L_{ij}(t)|\mathbf{u})$ are equal whenever $u_i = u_j$, that is, $(v|u) \leftrightarrow_{ij} (v'|u)$ iff $v = v'$ (providing, in the case $S = \Delta_n$, that $v < 1/2$, $u < 1/2$).

(ii) A projection rule with basic set R_+^n or Δ_n is *statistical* if it is regular, local, and the relation $(v|u) \leftrightarrow_{ij} (v'|u')$ is equivalent to $v/u = v'/u'$, with the additional conditions $v + v' < 1$, $u + u' < 1$ if $S = \Delta_n$.

The term "semisymmetric" refers to the fact that this weak and plausible postulate often implies the apparently much stronger property of symmetry (permutation invariance); see Theorem 2(i). It would not be unreasonable to use symmetry as a postulate, as did Shore and Johnson (1980), but by not doing so we will obtain stronger mathematical results with little additional effort.

Also the stronger postulate in (ii), applicable when $S = R_+^n$ or Δ_n, appears natural; it is particularly compelling in the latter case. Namely, if a prior guess about a probability mass function has to be updated subject to a single constraint that specifies the probability of a given set, it is standard to assign

probabilities to the elements of this set proportionally to the prior ones. Definition 4(ii) requires this for two-element sets only. The proportional updating, in general for constraints specifying the probabilities of several pairwise disjoint sets ["Jeffrey's rule," cf. Diaconis and Zabell (1982)] appears uncontroversial. It was also part of the axioms of Shore and Johnson (1980).

DEFINITION 5. (i) A projection rule with basic set $S = R^n$ or R^n_+ is *scale-invariant* if for every $\lambda > 0$, $L \in \mathscr{L}$ and $\mathbf{u} \in S$ we have $\Pi(\lambda L|\lambda\mathbf{u}) = \lambda\Pi(L|\mathbf{u})$.

(ii) A projection rule with basic set $S = R^n$ is *translation-invariant* if for every $L \in \mathscr{L}$, $\mathbf{u} \in S$ and $\mu \in R$, $\Pi(L + \mu\mathbf{1}|\mathbf{u} + \mu\mathbf{1}) = \Pi(L|\mathbf{u}) + \mu\mathbf{1}$.

Here λL and $L + \mu\mathbf{1}$ denote the set of vectors $\lambda\mathbf{v}$ and $\mathbf{v} + \mu\mathbf{1}$, respectively, such that $\mathbf{v} \in L$.

The intuitive meaning of these invariance properties is obvious, and they are clearly desirable. The characterization of all (regular and local) projection rules satisfying either or both of these postulates is not difficult but will be omitted because of its limited practical interest. Rather, these postulates will be used in connection with the next one only.

DEFINITION 6. (i) A projection rule is *subspace-transitive* if for arbitrary $L' \subset L$ in $\mathscr{L}$ and any $\mathbf{u} \in S$ we have

$$\Pi(L'|\mathbf{u}) = \Pi(L'|\Pi(L|\mathbf{u})). \tag{2.11}$$

(ii) A projection rule is *parallel-transitive* if (2.11) holds for subspaces L and L' that can be represented in the form $\{\mathbf{v}\colon \mathbf{A}\mathbf{v} = \mathbf{b}\}$ [cf. (2.2)] with the same matrix $\mathbf{A}$.

Subspace transitivity means that if updating a "prior guess" $\mathbf{u}$ based upon information specifying the feasible set L results in $\mathbf{v} = \Pi(L|\mathbf{u})$, and additional information leads to L' as the actual feasible set, then updating the "present guess" $\mathbf{v}$ on the basis of all available information leads to the same result as would the direct updating of the "prior guess" $\mathbf{u}$. For example, if an image reconstruction has been obtained from measurements $R_i f$, $i = 1, \ldots, k$, using some prior guess, and then further measurements $R_i f$, $i = k + 1, \ldots, k + l$, are made, the reconstruction from all the measurements $R_i f$, $i = 1, \ldots, k + l$, will be the same no matter whether the original prior guess or the previous reconstruction is used as "prior guess." Parallel transitivity has a similar intuitive meaning for the case when after having obtained the first reconstruction, the same measurements are repeated with results different from those before; the second reconstruction is now based on the new results, the previous contradicting ones being discarded.

Subspace transitivity is a highly desirable property also because it implies (and is actually equivalent to) the commutativity of two-stage updatings. Indeed, let us be given two sets of linear constraints determining subspaces L_1 and L_2 with $L_1 \cap L_2 \neq \varnothing$. Then (2.11) implies that updating a prior guess $\mathbf{u}$ based upon the first set of constraints and then updating the result $\Pi(L_1|\mathbf{u})$

based upon both sets of constraints, the result will be $\Pi(L_1 \cap L_2|\mathbf{u})$; the same result would be obtained if the first updating were based on the second set of constraints. Of course, this commutativity holds for "proper" two-stage updatings only, that is, when the constraints used in the first stage are not discarded in the second stage.

For regular projection rules, parallel transitivity implies subspace transitivity, and more generally, also that (2.11) holds whenever $L = \{\mathbf{v}\colon \mathbf{Av} = \mathbf{b}\}$, $L' = \{\mathbf{v}\colon \mathbf{A}'\mathbf{v} = \mathbf{b}'\}$, where the matrix $\mathbf{A}'$ contains all rows of $\mathbf{A}$. In fact, then $\tilde{L} = \{\mathbf{v}\colon \mathbf{A}'\mathbf{v} = \mathbf{A}'\mathbf{v}^*\} \subset L$, where $\mathbf{v}^* = \Pi(L|\mathbf{u})$, and thus $\Pi(\tilde{L}|\mathbf{u}) = \mathbf{v}^*$ by the consistency axiom. Then parallel transitivity implies that

$$\Pi(L'|\mathbf{u}) = \Pi\big(L'|\Pi(\tilde{L}|\mathbf{u})\big),$$

and (2.11) holds as claimed.

We will show that for regular, local projection rules the two kinds of transitivity defined above are actually equivalent. The family of all regular, local and transitive projection rules will be characterized in Theorem 3.

Combining results characterizing projection rules with properties stated in Definitions 4–6 will lead us to the least squares and I-divergence projection rules as the only ones that simultaneously satisfy some intuitively appealing postulates. On the other hand, a single postulate (in addition to regularity and locality) will also suffice to uniquely characterize these projection rules, as well as the least squares and I-divergence selection rules. To formulate this postulate (Definition 7 below), it is necessary to consider vectors indexed by pairs of integers (i, j), $1 \le i \le m$, $1 \le j \le n$, rather than by integers $1 \le i \le n$. This is the natural representation for vectors describing two-dimensional objects, such as images.

The *marginals* of $\mathbf{v} = \{v_{ij}\colon 1 \le i \le m,\ 1 \le j \le n\} \in R^{mn}$ are the vectors $\overline{\mathbf{v}} = (\overline{v}_1, \dots, \overline{v}_m)^T \in R^m$, $\overline{\overline{\mathbf{v}}} = (\overline{\overline{v}}_1, \dots, \overline{\overline{v}}_n)^T \in R^n$, where

$$\overline{v}_i = \sum_{j=1}^{n} v_{ij}, \qquad \overline{\overline{v}}_j = \sum_{i=1}^{m} v_{ij}. \tag{2.12}$$

As before, we consider three choices of our basic set S, namely, $S = R^{mn}$ or R_+^{mn} or Δ_{mn}.

For any $\mathbf{v} = \{v_{ij}\} \in S$, we denote by $L_{\mathbf{v}}$ the subspace of S consisting of those vectors that have the same marginals as $\mathbf{v}$, that is,

$$L_{\mathbf{v}} = \big\{\mathbf{w}\colon \overline{\mathbf{w}} = \overline{\mathbf{v}};\ \overline{\overline{\mathbf{w}}} = \overline{\overline{\mathbf{v}}}\big\}. \tag{2.13}$$

We say that $\mathbf{v} = \{v_{ij}\}$ is of *sum* or *product form* if

$$v_{ij} = s_i + t_j \quad \text{or} \quad v_{ij} = s_i t_j, \tag{2.14}$$

respectively.

DEFINITION 7. A selection rule Π with basic set S as above is *composition-consistent*, more precisely, *sum-* or *product-consistent*, if

$\Pi(L_{\mathbf{v}}) = \mathbf{v}$ whenever $\mathbf{v} \in S$ is of sum or product form, respectively. A projection rule is *composition-consistent* (sum- or product-consistent) if its component selection rules $\Pi(\cdot|\mathbf{u})$ have this property whenever $\mathbf{u}$ itself is of sum or product form, respectively.

Intuitively, the vectors $\mathbf{v} = \{v_{ij},\ 1 \le i \le m,\ 1 \le j \le n\}$ are interpreted as compositions of two interacting components, the individual components being represented by the marginals $\bar{\mathbf{v}}$ and $\bar{\bar{\mathbf{v}}}$. Then $L_{\mathbf{v}}$ being the feasible set means that the available information specifies the individual components but nothing else. The postulate of composition consistency formalizes the intuitive requirement that on the basis of such information we should infer "no interaction," unless a prior guess is available that implies interaction. Implicit in this interpretation is that "no interaction" is represented by the sum or product form of $\mathbf{v}$. In particular, for inferring probability mass functions, product consistency is a hardly avoidable postulate. It could be argued, though with less weight, that product consistency is an appropriate axiom also for image reconstruction, that is, that if the marginals of an image were known and nothing else (an unlikely situation in practice), the "best" reconstruction would be of product form. In inverse problems without a positivity constraint, the sum form of $\mathbf{v}$ appears to be the natural description of "no interaction," suggesting that in this case the postulate of sum consistency should be adopted.

REMARKS. (i) Product-consistent selection rules cannot exist for $S = R^{mn}$ because in that case different elements of S, each of product form, can have the same marginals. On the other hand, sum-consistent selection rules, satisfying the continuity axiom in Definition 2, cannot exist for $S = R_+^{mn}$ or $S = \Delta_{mn}$. To see this, consider a sequence $\mathbf{v}^{(k)}$ of elements of S of sum form $v_{ij}^{(k)} = s_i^{(k)} + t_j^{(k)}$, $s_i^{(k)} > 0$, $t_j^{(k)} > 0$, such that $s_i^{(k)} \to s_i$, $t_j^{(k)} \to t_j$, where $s_1 = t_1 = 0$ and $s_i > 0$ for $i > 1$, $t_j > 0$ for $j > 1$. Then $\mathbf{v}^{(k)} \to \mathbf{v}$ and $L_{\mathbf{v}^{(k)}} \to L_{\mathbf{v}}$, where $v_{ij} = s_i + t_j$. Thus for a sum-consistent selection rule we should have

$$\Pi(L_{\mathbf{v}}) = \lim_{k\to\infty} \Pi(L_{\mathbf{v}^{(k)}}) = \lim_{k\to\infty} \mathbf{v}^{(k)} = \mathbf{v},$$

a contradiction because $\mathbf{v} \notin S$.

(ii) In the case $S = R^{mn}$, every $L_{\mathbf{v}}$ as in (2.13) contains an element of sum form; hence for a sum-consistent selection rule Π, $\Pi(L_{\mathbf{v}})$ is always of sum form. It follows, subject to regularity, that $\mathbf{v}^0 = \mathbf{v}^0(\Pi)$ is of sum form, because $\mathbf{v}^0 = \Pi(L_{\mathbf{v}^0})$ by the consistency axiom. Similarly, in the cases $S = R_+^{mn}$ or Δ_{mn}, if Π is a product-consistent (regular) selection rule, then $\mathbf{v}^0(\Pi)$ is of product form.

(iii) For the case $S = \Delta_{mn}$, the postulate of product consistency is similar to but weaker than Shore and Johnson's (1980) "system independence" postulate.

3. Statement of results. Recall that our basic set S is either R^n, R_+^n or Δ_n, the family of affine subspaces of S is denoted by $\mathscr{L}$, and V denotes R, R_+ or the interval $(0, 1)$, according to the choice of S. Throughout this section, we assume that $n \geq 3$ (if $S = R^n$ or R_+^n) or $n \geq 5$ (if $S = \Delta_n$).

The following terminology will facilitate the statement of our results.

DEFINITION 8. A n-tuple of functions $f_1, \ldots, f_n$ defined on V will be called a *standard n-tuple* with 0 at $\mathbf{v}^0 = (v_1^0, \ldots, v_n^0)^T \in S$ if:

(i) f_i is continuously differentiable and vanishes together with its derivative at v_i^0, $i = 1, \ldots, n$,

(ii) in the cases $S = R_+^n$ or Δ_n, $\lim_{v \to 0} f_i'(v) = -\infty$, $i = 1, \ldots, n$,

(iii) the function

$$F(\mathbf{v}) = \sum_{i=1}^{n} f_i(v_i) \tag{3.1}$$

is nonnegative and strictly quasiconvex on S, that is, for any $\mathbf{v}$ and $\mathbf{v}'$ in S and any $0 < \alpha < 1$ we have

$$F(\alpha\mathbf{v} + (1 - \alpha)\mathbf{v}') < \max(F(\mathbf{v}), F(\mathbf{v}')). \tag{3.2}$$

REMARK. The functions f_i in a standard n-tuple must be convex if $S = R^n$ or R_+^n but not necessarily if $S = \Delta_n$. The proof of this is omitted in order to save space.

THEOREM 1. (i) *For any regular, local selection rule* Π: $\mathscr{L} \to S$, *there exists a standard n-tuple* $f_1, \ldots, f_n$ *with* 0 *at* $\mathbf{v}^0 = \mathbf{v}^0(\Pi)$ *such that* Π *is generated by the function* F *in* (3.1). *Conversely, any* F *as in* (3.1) *generates a regular, local selection rule* Π *with* $\mathbf{v}^0(\Pi) = \mathbf{v}^0$.

(ii) *For any regular, local projection rule, there exist functions* $f_i(v|u)$, $u \in V$, $v \in V$, *such that for every fixed* $\mathbf{u} = (u_1, \ldots, u_n)^T \in S$, *the functions* $f_i(v|u_i)$ *form a standard n-tuple with* 0 *at* $\mathbf{u}$ *and the given projection rule is generated by*

$$F(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^{n} f_i(v_i|u_i). \tag{3.3}$$

Conversely, any such F *generates a regular, local projection rule.*

(iii) *Two functions* $F(\mathbf{v}) = \sum_{i=1}^n f_i(v_i)$, $\tilde{F}(\mathbf{v}) = \sum_{i=1}^n \tilde{f}_i(v_i)$ *or* $F(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^n f_i(v_i|u_i)$, $\tilde{F}(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^n \tilde{f}_i(v_i|u_i)$ *as in* (i) *or* (ii) *generate the same selection or projection rule, respectively, if and only if* $\tilde{f}_i = cf_i$, $i = 1, \ldots, n$, *for some constant* $c > 0$.

REMARK. In part (iii), the assumption that F and $\tilde{F}$ be of the stated form is essential. Otherwise, $F(\mathbf{v})$ and $F = \Phi(F(\mathbf{v}))$ generate the same selection rule for any strictly increasing function Φ, and $F(\mathbf{v}|\mathbf{u})$ and $\tilde{F} = \Phi(F(\mathbf{v}|\mathbf{u}), \mathbf{u})$ generate the same projection rule whenever $\Phi(\cdot|\mathbf{u})$ is strictly increasing for every fixed $\mathbf{u} \in S$.

THEOREM 2. (i) *A projection rule with basic set* $S = R^n$ *or* R_+^n *is regular, local and semisymmetric* (*Definition* 4) *if and only if it is generated by* $F(\mathbf{v}|\mathbf{u})$ *as in Theorem* 1(ii) *with* f_i *not depending on* i.

(ii) *A projection rule with basic set* $S = R_+^n$ *or* Δ_n *is statistical* (*Definition* 4) *if and only if it is generated by*

$$F(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^{n} u_i f\left(\frac{v_i}{u_i}\right), \tag{3.4}$$

for some continuously differentiable, strictly convex function f *on* R_+, *with* $f(1) = f'(1) = 0$ *and* $\lim_{t \to 0} f'(t) = -\infty$.

Functions of probability distributions of form (3.4), as measures of distance motivated by information theory, were introduced by Csiszár (1963) under the name f-divergences [cf. also Csiszár (1967)] and independently by Ali and Silvey (1966). For their applications in statistics, see, for example, Liese and Vajda (1987). The conditions on the derivative of f are not part of the original definition of f-divergences. Notice that the condition $f'(1) = 0$ is essential only in the case $S = R_+^n$ because if $S = \Delta_n$, then any $f(t)$ and $\tilde{f}(t) = f(t) + c(t - 1)$ define the same F in (3.4). The condition $\lim_{t \to 0} f'(t) = -\infty$ is necessary for F in (3.4) to generate a projection rule, that is, to ensure that its minimum in $\mathbf{v}$ be attained on every $L \in \mathscr{L}$.

THEOREM 3. (i) *For any regular, local and subspace-transitive projection rule* (*Definition* 6), *there exists a standard* n*-tuple* $\varphi_1, \ldots, \varphi_n$ *such that* $\Phi(\mathbf{v}) = \sum_{i=1}^{n} \varphi_i(v_i)$ *is strictly convex on* S *and the given projection rule is generated by*

$$\begin{aligned} F(\mathbf{v}|\mathbf{u}) &= \Phi(\mathbf{v}) - \Phi(\mathbf{u}) - (\operatorname{grad} \Phi(\mathbf{u}))^T(\mathbf{v} - \mathbf{u}) \\ &= \sum_{i=1}^{n} \left(\varphi_i(v_i) - \varphi_i(u_i) - \varphi_i'(u_i)(v_i - u_i)\right). \end{aligned} \tag{3.5}$$

(ii) *Any* F *as in part* (i) *generates a regular, local and parallel-transitive projection rule.*

On account of this theorem, in the sequel we need not distinguish between the two kinds of transitivity.

COROLLARY. *The only transitive statistical projection rule* (*with* $S = R_+^n$ *or* Δ_n) *is the I-divergence projection rule* (*cf. Example* 2).

The class of measures of distance associated as in Theorem 3 with strictly convex functions Φ (not necessarily of a sum form) was introduced by Bregman (1967). He developed an iterative algorithm for minimizing Φ under linear (and, more generally, under linear inequality) constraints, whose steps involved projections in the sense of the distance corresponding to Φ [cf. also Censor and Lent (1981)].

Bregman's divergences include squared Euclidean distance and I-divergence, and share with these the property

$$F(\mathbf{v}|\mathbf{u}) + F(\mathbf{w}|\mathbf{v}) = F(\mathbf{w}|\mathbf{u}) \quad \text{if } \Pi(L|\mathbf{u}) = \mathbf{v}, \qquad \mathbf{w} \in L. \tag{3.6}$$

Notice that for squared Euclidean distance, (3.6) is just the Pythagorean theorem. The fact that I-divergence also has this Pythagorean property plays a key role in its applications in statistics [cf. Kullback (1959)].

A result related to but not directly comparable with Theorem 3 was recently obtained by Jones and Byrne (1990), generalizing the previous results of Jones (1989). They showed that among the continuous analogs of the distances of form (3.3), only the analogs of those in Theorem 3 satisfied a postulate called projectivity. That postulate is strongly motivated from the point of view of inference, but it also involves the distance itself, as opposed to transitivity which is a property of the projection rule alone. Jones and Byrne (1990) pointed out that their distances satisfied (3.6), and, in fact, that (3.6) was equivalent to the projectivity postulate.

THEOREM 4. (i) *A projection rule with basic set $S = R^n$ is regular, local, transitive and both location and scale-invariant if and only if it is a weighted least squares projection rule (cf. Example 1). The above properties and semisymmetry uniquely characterize the least squares projection rule.*

(ii) *A projection rule with basic set $S = R_+^n$ is regular, local, transitive and scale-invariant if and only if it is generated by $F(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^n a_i h_\alpha(v_i|u_i)$, where $a_1, \ldots, a_n$ are positive constants, $\alpha \le 1$, and*

$$h_\alpha(v|u) = \begin{cases} v \log \dfrac{v}{u} - v + u, & \text{if } \alpha = 1, \\ \log \dfrac{u}{v} + \dfrac{v}{u} - 1, & \text{if } \alpha = 0, \\ \dfrac{1}{\alpha}(u^\alpha - v^\alpha) + u^{\alpha-1}(v - u), & \text{if } 0 < \alpha < 1 \text{ or } \alpha < 0. \end{cases} \tag{3.7}$$

A projection rule with the above properties is also semisymmetric if and only if it is generated by

$$F_\alpha(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^n h_\alpha(v_i|u_i), \qquad \alpha \le 1. \tag{3.8}$$

REMARK. The one-parameter family (3.8) contains two well-known distances: I-divergence ($\alpha = 1$) and Itakura and Saito (1968) distance ($\alpha = 0$). Recent results of Jones and Trutzer (1989) indicate that (continuous versions of) the members of this family with $\alpha = 1/m$ (m some integer) may be preferable to the Itakura–Saito distance in spectrum reconstruction, a traditional field of application of the latter.

Our last result provides a characterization of the least squares and I-divergence selection and projection rules that, instead of invariance or transitivity, rely on composition consistency (Definition 7). We emphasize that this characterization also applies to individual selection rules, rather than to projection rules only. On the other hand, we have to assume that the basic set S is as described before Definition 7. Intuitively, the inference problem in question should relate to objects with (at least) two components.

THEOREM 5. *Let S be as in Definition 7, $S = R^{mn}$ or R_+^{mn} or Δ_{mn}, with $m \geq 2$, $n \geq 2$, and if $S = \Delta_{mn}$, $m + n \geq 5$.*

(i) *In the case $S = R^{mn}$, the regular, local, sum-consistent selection and projection rules are exactly those least squares selection rules for which $\mathbf{v}^0$ is of sum form, and the least squares projection rule, respectively (cf. Example 1).*

(ii) *In the cases $S = R_+^{mn}$ or Δ_{mn}, the regular, local and product-consistent selection and projection rules are exactly those I-divergence selection rules for which $\mathbf{v}^0$ is of product form, and the I-divergence projection rule, respectively (cf. Example 2).*

COROLLARY. (i) *For $S = R^{mn}$, the standard least squares selection rule is the unique regular, local, sum-consistent selection rule for which $\mathbf{v}^0(\Pi) = \mathbf{0}$.*

(ii) *For $S = \Delta_{mn}$, the maximum entropy selection rule is the unique regular, local, product-consistent selection rule for which $\mathbf{v}^0(\Pi)$ is the "uniform distribution" $(1/mn)\mathbf{1}$. The same holds also for $S = R_+^{mn}$ if the last condition is replaced by $\mathbf{v}^0(\Pi) = (1/e)\mathbf{1}$.*

4. Basic lemmas.

LEMMA 1. *For a regular selection rule Π: $S \to \mathscr{L}$, for every $L' \in \mathscr{L}$ of dimension less than $n - 1$ (if $S = R^n$ or R_+^n) or less than $n - 2$ (if $S = \Delta_n$), there exists $L \in \mathscr{M}$ [cf. (2.3)] such that $L' \subset L$ and $\Pi(L') = \Pi(L)$.*

COROLLARY . *For a regular Π and $L' \in \mathscr{L}$, $L \in \mathscr{M}$, the equality $\Pi(L') = \Pi(L)$ implies $L' \subset L$ unless $\Pi(L') = \mathbf{v}^0(\Pi)$.*

PROOF. Clearly, it suffices to prove that if $\dim L' = d$ with $0 \leq d < n - 1$ (or $0 \leq d < n - 2$), then there exists an $L \in \mathscr{L}$ such that $L \supset L'$, $\dim L = d + 1$, $\Pi(L') = \Pi(L)$. Further, instead of the last equality, it suffices to show that $\Pi(L) \in L'$ because this, by the consistency axiom in Definition 2, already implies $\Pi(L') = \Pi(L)$.

Now, pick any $(d + 1)$-dimensional $L_1 \supset L'$ and suppose that $\Pi(L_1) \notin L'$. Then we will "rotate" L_1 to obtain a family $\{L_t\colon 0 \leq t \leq 2\}$ and show that $\Pi(L_t) \in L'$ for some t. To this end, pick any $\mathbf{v}_0 \notin L_1$ in S, set $\mathbf{v}_1 = \Pi(L_1)$ and let $\mathbf{v}_2$ be such that some interior point of the segment $[\mathbf{v}_0, \mathbf{v}_2]$ is in L'. Set $\mathbf{v}_t = (1 - t)\mathbf{v}_0 + t\mathbf{v}_1$ if $0 \leq t \leq 1$ and $\mathbf{v}_t = (2 - t)\mathbf{v}_1 + (t - 1)\mathbf{v}_2$ if $1 \leq t \leq 2$,

and let L_t denote the subspace of S spanned by L' and $\mathbf{v}_t$. Then $L_0 = L_2$ and $L_{t_1} \cap L_{t_2} = L'$ if $0 \le t_1 < t_2 < 2$.

By the continuity axiom, $\{\Pi(L_t)\colon 0 \le t \le 2\}$ is a continuous closed curve in the subspace $\tilde{L}$ spanned by L_1 and $\mathbf{v}_0$ ($t = 0$ and $t = 2$ representing the same point). For $\varepsilon > 0$ sufficiently small, $\Pi(L_{1-\varepsilon})$ and $\Pi(L_{1+\varepsilon})$—which are arbitrarily close to $\Pi(L_1) = \mathbf{v}_1$—are separated by L_1 within $\tilde{L}$; hence there exists some t with $|t - 1| > \varepsilon$ for which $\Pi(L_t) \in L_1$. Then $\Pi(L_t) \in L_t \cap L_1 = L'$, and the proof of Lemma 1 is complete.

The corollary is immediate, because for $\tilde{L} \in \mathscr{M}$ containing L' such that $\Pi(L') = \Pi(\tilde{L})$, the distinctness axiom implies $\tilde{L} = L$. □

LEMMA 2. *The restriction of a regular selection rule* $\Pi\colon \mathscr{L} \to S$ *to* $\mathscr{M} \setminus \mathscr{L}^0$ *is a homeomorphism onto* $S \setminus \{\mathbf{v}^0\}$, *where* $\mathbf{v}^0 = \mathbf{v}^0(\Pi)$ *and* $\mathscr{L}^0 = \{L\colon \mathbf{v}^0 \in L\}$.

PROOF. Applying Lemma 1 to the zero-dimensional subspace $L' = \{\mathbf{v}\}$, it follows that for each $\mathbf{v} \in S$ there exists $L \in \mathscr{M}$ such that $\Pi(L) = \mathbf{v}$. If $\mathbf{v} \ne \mathbf{v}^0$, then $L \notin \mathscr{L}^0$, by the consistency axiom. Thus Π maps $\mathscr{M} \setminus \mathscr{L}^0$ onto $S \setminus \{\mathbf{v}^0\}$. This mapping is one-to-one and continuous by the distinctness and continuity axioms. It remains to prove the continuity of the inverse. In other words, we have to show that $\Pi(L_k) \to \Pi(L) \ne \mathbf{v}^0$ implies $L_k \to L$.

We do this by showing that every subsequence of $\{L_k\}$ contains a subsequence converging to L. Write

$$L_k = \begin{cases} \{\mathbf{v}\colon \mathbf{a}_k^T\mathbf{v} = b_k\}, & \text{if } S = R^n \text{ or } S = R_+^n, \\ \{\mathbf{v}\colon \mathbf{a}_k^T\mathbf{v} = b_k;\ \mathbf{1}^T\mathbf{v} = 1\}, & \text{if } S = \Delta_n \end{cases}$$

[cf. (2.3)]. Here we may suppose that $\|\mathbf{a}_k\| = 1$ and in the case $S = \Delta_n$ also that $\mathbf{a}_k \perp \mathbf{1}$.

Now, any subsequence of $\{L_k\}$ contains a subsequence $\{L_{k_i}\}$ such that $\mathbf{a}_{k_i} \to \tilde{\mathbf{a}}$, say. Write $\Pi(L_k) = \mathbf{v}_k$, $\Pi(L) = \mathbf{v}^*$, $\tilde{\mathbf{a}}^T\mathbf{v}^* = \tilde{b}$. As $\mathbf{v}_k \to \mathbf{v}^*$ by assumption, $\mathbf{a}_{k_i} \to \tilde{\mathbf{a}}$ implies that $b_{k_i} = \mathbf{a}_{k_i}^T\mathbf{v}_k \to \tilde{\mathbf{a}}^T\mathbf{v}^* = \tilde{b}$. This means that if the set

$$\tilde{L} = \begin{cases} \{\mathbf{v}\colon \tilde{\mathbf{a}}^T\mathbf{v} = \tilde{b}\}, & \text{if } S = R^n \text{ or } R_+^n, \\ \{\mathbf{v}\colon \tilde{\mathbf{a}}^T\mathbf{v} = \tilde{b},\ \mathbf{1}^T\mathbf{v} = 1\}, & \text{if } S = \Delta_n, \end{cases}$$

is in $\mathscr{M}$, we have $L_{k_i} \to \tilde{L}$. But $\tilde{L} \in \mathscr{M}$ holds because (i) $\tilde{L} \ne \varnothing$ (namely, $\mathbf{v}^* \in \tilde{L}$ by the definition of $\tilde{b}$) and (ii) $\|\tilde{\mathbf{a}}\| = 1$ and, if $S = \Delta_n$, also $\tilde{\mathbf{a}} \perp \mathbf{1}$.

We have proved that every subsequence of $\{L_k\}$ contains a convergent subsequence $L_{k_i} \to \tilde{L}$. By the continuity axiom, here $\Pi(\tilde{L}) = \lim_{i\to\infty} \Pi(L_{k_i}) = \Pi(L) \ne \mathbf{v}^0$, and hence necessarily $\tilde{L} = L$ by the distinctness axiom. This completes the proof of Lemma 2. □

We will need the following notation, for $k \le n$: The vectors in R^k whose components are all 0 or all 1, will be denoted by $\mathbf{0}_k$ and $\mathbf{1}_k$, respectively (thus

$\mathbf{0} = \mathbf{0}_n$, $\mathbf{1} = \mathbf{1}_n$). Two vectors in R^k will be called *equivalent*, denoted by $\mathbf{a} \sim \tilde{\mathbf{a}}$, iff for some $\lambda \neq 0$,

$$\mathbf{a} = \begin{cases} \lambda\tilde{\mathbf{a}}, & \text{if } S = R^n \text{ or } S = R_+^n, \\ \lambda\tilde{\mathbf{a}} + \mu\mathbf{1}_k,\ \mu \in R, & \text{if } S = \Delta_n. \end{cases}$$

Observe that in the representation

$$L = \begin{cases} \{\mathbf{v}\colon \mathbf{a}^T\mathbf{v} = b\}, & \text{if } S = R^n \text{ or } R_+^n, \\ \{\mathbf{v}\colon \mathbf{a}^T\mathbf{v} = b,\ \mathbf{1}^T\mathbf{v} = 1\}, & \text{if } S = \Delta_n, \end{cases} \tag{4.1}$$

of a subspace $L \in \mathscr{M}$, the vector $\mathbf{a} \in R^n$ is determined up to equivalence, and it is not equivalent to $\mathbf{0}$.

LEMMA 3. *Let* $\Pi\colon \mathscr{L} \to S$ *be a regular, local selection rule, and let* $\leftrightarrow_{ij}$ *be the relation introduced in the passage containing* (2.10).

(i) *Let* $L \in \mathscr{M}$, $\Pi(L) = \mathbf{v}^* \neq \mathbf{v}^0 = \mathbf{v}^0(\Pi)$. *Then* $v_i^* \leftrightarrow_{ij} v_j^*$ *if and only if* $a_i = a_j$ *in the representation* (4.1) *of* L; *further, in the cases* $S = R^n$ *or* $S = R_+^n$, $v_i^* = v_i^0$ *if and only if* $a_i = 0$.

(ii) *Let* L *and* $\tilde{L}$ *be both in* $\mathscr{M}$ *with* $\Pi(L) \neq \mathbf{v}^0$, $\Pi(\tilde{L}) \neq \mathbf{v}^0$, *and let* $J \subset \{1, \ldots, n\}$. *Then* $\Pi_J(L) = \Pi_J(\tilde{L})$ *implies* $\mathbf{a}_J \sim \tilde{\mathbf{a}}_J$ *for* $\mathbf{a}$ *and* $\tilde{\mathbf{a}}$ *representing* L *and* $\tilde{L}$ *as in* (4.1). *In other words,* $\mathbf{a}_J$ *is determined by* $\Pi_J(L)$ *up to equivalence.*

COROLLARY. *If* $v_i \leftrightarrow_{ij} v_j$ *and* $v_j \leftrightarrow_{jk} v_k$ *for some* $\{i, j, k\} \subset \{1, \ldots, n\}$, *then also* $v_i \leftrightarrow_{ik} v_k$, *provided in the case* $S = \Delta_n$ *that* $v_i + v_j + v_k < 1$.

PROOF. We will show that for L as in (i), arbitrary $J = \{j_1, \ldots, j_k\} \subset \{1, \ldots, n\}$, $\hat{\mathbf{a}} \in R^k$, and for

$$L' = \begin{cases} \{\mathbf{v}\colon \hat{\mathbf{a}}^T\mathbf{v}_J = \hat{\mathbf{a}}^T\mathbf{v}_J^*\}, & \text{if } S = R^n \text{ or } R_+^n, \\ \{\mathbf{v}\colon \hat{\mathbf{a}}^T\mathbf{v}_J = \hat{\mathbf{a}}^T\mathbf{v}_J^*,\ \mathbf{1}_k^T\mathbf{v}_J = \mathbf{1}_k^T\mathbf{v}_J^*,\ \mathbf{1}_{n-k}^T\mathbf{v}_{J^c} = \mathbf{1}_{n-k}^T\mathbf{v}_{J^c}^*\}, & \\ & \text{if } S = \Delta_n, \end{cases} \tag{4.2}$$

the following holds: $\mathbf{a}_J \sim \hat{\mathbf{a}}$ is a sufficient condition for

$$\Pi_J(L) = \Pi_J(L') \tag{4.3}$$

and this condition is also necessary unless $\mathbf{a}_J \sim \mathbf{0}_k$.

To prove this, set

$$L'' = L' \cap \{\mathbf{v}\colon \mathbf{v}_{J^c} = \mathbf{v}_{J^c}^*\}. \tag{4.4}$$

Then, by locality, $\Pi_J(L') = \Pi_J(L'')$. Since, of course, $\Pi_{J^c}(L'') = \mathbf{v}_{J^c}^*$, it follows that (4.3) holds if and only if $\Pi(L'') = \mathbf{v}^*$.

Now, if $\mathbf{a}_J \sim \hat{\mathbf{a}}$, then $L'' \subset L$. By the consistency axiom, the latter implies $\Pi(L'') = \mathbf{v}^*$. Thus $\mathbf{a}_J \sim \hat{\mathbf{a}}$ is sufficient for (4.3), as claimed.

Conversely, if (4.3) and therefore $\Pi(L'') = \mathbf{v}^*$ hold, the corollary of Lemma 1 yields $L'' \subset L$. Comparing (4.1), (4.3) and (4.4), $L'' \subset L$ means that $\hat{\mathbf{a}}^T \mathbf{v}_J = \hat{\mathbf{a}}^T \mathbf{v}_J^*$ implies $\mathbf{a}_J^T \mathbf{v}_J = \mathbf{a}_J^T \mathbf{v}_J^*$ (subject to the additional constraint $\sum_{j \in J} v_j = \sum_{j \in J} v_j^*$ if $S = \Delta_n$). Clearly, this implication holds iff the two conditions represent the same constraint, except for the case $\mathbf{a}_J \sim \mathbf{0}_k$. This proves that $\mathbf{a}_J \sim \hat{\mathbf{a}}$ is indeed necessary for (4.3) unless $\mathbf{a}_J \sim \mathbf{0}_k$.

To prove assertion (i) of the lemma, suppose first that $S = R^n$ or $S = R_+^n$ and apply the result just proved to $J = \{i\}$, $\hat{a} = 0$. Then L' in (4.2) equals S and (4.3) means that $v_i^* = v_i^0$. Thus we obtain that $a_i = 0$ implies $v_i^* = v_i^0$ and conversely, if $a_i \neq 0$, then $v_i^* = v_i^0$ cannot hold. Next, apply our result to $J = \{i, j\}$, $\hat{\mathbf{a}} = (1, 1)^T$. Then L' in (4.2) equals the special subspace $L_{ij}(t)$ (with $t = v_i^* + v_j^*$) that defines the relation $\leftrightarrow_{ij}$ [cf. (2.10)]. Hence (4.3) is equivalent to $v_i^* \leftrightarrow_{ij} v_j^*$. Now if $a_i = a_j$, then $\mathbf{a}_J \sim \hat{\mathbf{a}} = (1, 1)^T$, except when $S = R^n$ or $S = R_+^n$ and $a_i = a_j = 0$. Thus we have that $a_i = a_j$ implies $v_i^* \leftrightarrow_{ij} v_j^*$, because the mentioned exceptional case has already been covered.

Conversely, if $a_i \neq a_j$, then neither $\mathbf{a}_J \sim \mathbf{0}_2$ nor $\mathbf{a}_J \sim \hat{\mathbf{a}}$; hence $v_i^* \leftrightarrow_{ij} v_j^*$ cannot hold.

To prove assertion (ii), suppose that $\Pi_J(L) = \Pi_J(\tilde{L}) = \mathbf{v}_J^*$ and consider L' as in (4.2) with $\hat{\mathbf{a}} = \tilde{\mathbf{a}}_J$. Then $\Pi_J(L) = \Pi_J(\tilde{L}) = \Pi_J(L')$, by assumption and the sufficient condition for (4.3). This implies, by the necessary condition, that $\mathbf{a}_J \sim \tilde{\mathbf{a}}_J$, except perhaps when $\mathbf{a}_J \sim \mathbf{0}_k$. Since the roles of L and $\tilde{L}$ are symmetric, we similarly have $\mathbf{a}_J \sim \tilde{\mathbf{a}}_J$ when $\mathbf{a}_J$ is not equivalent to $\mathbf{0}_k$, whereas the same trivially holds when both $\mathbf{a}_J \sim \mathbf{0}_k$ and $\tilde{\mathbf{a}}_J \sim \mathbf{0}_k$.

To prove the corollary, pick any $\mathbf{v}^* \in S$ with $v_i^* = v_i$, $v_j^* = v_j$, $v_k^* = v_k$. If $\mathbf{v}^* \neq \mathbf{v}^0$, the hypotheses $v_i \leftrightarrow_{ij} v_j$ and $v_j \leftrightarrow_{jk} v_k$ imply by Lemmas 2 and 3(i) that $\mathbf{v}^* = \Pi(L)$ for some L as in (4.1) with $a_i = a_j = a_k$. This, in turn, implies [again, by Lemma 3(i)] that $v_i \leftrightarrow_{ik} v_k$, whereas the latter is obvious (by the consistency axiom) if $\mathbf{v}^* = \mathbf{v}^0$. □

LEMMA 4. (i) *Let $F_{ij}(u, v)$, $i \neq j$, $\{i, j\} \subset \{1, \dots, n\}$, $n \geq 3$, be real-valued functions defined for $u \in A_i$, $v \in A_j$, where $A_1, \dots, A_n$ are arbitrary sets. If these F_{ij} satisfy the functional equations*

$$F_{ij}(u, v) F_{jk}(v, w) = F_{ik}(u, w), \qquad F_{ij}(u, v) F_{ji}(v, u) = 1, \tag{4.5}$$

then there exist functions g_i defined on A_i, $i = 1, \dots, n$, such that for every i and j,

$$F_{ij}(u, v) = \frac{g_i(u)}{g_i(v)}. \tag{4.6}$$

(ii) *Let $n \geq 4$ and let $B_{ij} \subset A_i \times A_j$, $C_{ijk} \subset A_i \times A_j \times A_k$ be sets such that (1) $(u, v) \in B_{ij}$ if and only if $(v, u) \in B_{ji}$; (2) $(u, v, w) \subset C_{ijk}$ implies that $(u, v) \in B_{ij}$, $(u, w) \in B_{ik}$, $(v, w) \in B_{jk}$; (3) for any distinct i, j, k, l and any $u \in A_i$, $v \in A_j$, $w \in A_k$ there exists $s \in A_l$ with $(u, s) \in B_{il}$, $(v, s) \in B_{jl}$, $(w, s) \in B_{kl}$; and (4) for any distinct i, j, k, l and any u, v, s, s' with $(u, s), (u, s')$ in B_{il} and $(v, s), (v, s')$ in B_{jl}, there exists $w \in A_k$ with*

$(u, w, s), (u, w, s')$ in C_{ikl} and $(s, w, v), (s', w, v)$ in C_{lkj}. Then if the functions F_{ij} are defined on the sets B_{ij} and the equations (4.5) *hold for $(u, v, w) \in C_{ijk}$, the conclusion of part* (i) *still holds.*

PROOF. (i) Fix some k and $\overline{w} \in A_k$, and write $g_i(u) = F_{ik}(u, \overline{w})$ for $i \neq k$. Then $g_i(u) \neq 0$ by the second part of (4.5), and the first part of (4.5) gives (4.6) if i, j are both different from k. In addition, (4.5) implies that

$$\frac{g_j(v)}{F_{jk}(v, w)} = \frac{g_i(u)}{F_{ik}(u, w)},$$

for every $i \neq k$, $j \neq k$, $u \in A_i$, $v \in A_j$. Denoting the common value of these quotients by $g_k(w)$, we obtain (4.6) also for $j = k$. Finally, the validity of (4.6) for the remaining case $i = k$ follows from the second equation in (4.5).

(ii) On account of (i) it suffices to prove that the functions F_{ij} can be extended from B_{ij} to $A_i \times A_j$ so that the equations (4.5) remain valid. To this end, given $u \in A_i$, $v \in A_j$, pick arbitrarily $l \neq k$ (both different from i, j) and s and s' in A_l such as in hypothesis (4); choose $w \in A_k$ according to that hypothesis. Then, applying the first and then the second part of (4.5), we get

$$\begin{aligned} F_{il}(u, s)F_{lj}(s, v) &= F_{ik}(u, w)F_{kl}(w, s)F_{lk}(s, w)F_{kj}(w, v) \\ &= F_{ik}(u, w)F_{kj}(w, v) \end{aligned}$$

and similarly

$$F_{il}(u, s')F_{lj}(s', v) = F_{ik}(u, w)F_{kj}(w, v).$$

This means that

$$\tilde{F}_{ij}(u, v) = F_{il}(u, s)F_{lj}(s, v) \tag{4.7}$$

is well defined because the right-hand side does not depend on l and s [subject to $(u, s) \in B_{il}$, $(v, s) \in B_{jl}$, the latter being equivalent to $(s, v) \in B_{lj}$]. Clearly, (4.7) defines an extension of F_{ij} to $A_i \times A_j$. To see that these extensions $\tilde{F}_{ij}$ satisfy the functional equations (4.5) for every $u \in A_i$, $v \in A_j$, $w \in A_k$, let l be different from i, j, k and pick $s \in A_l$ according to hypothesis (3). Then by (4.7) and the second part of (4.5)

$$\begin{aligned} \tilde{F}_{ij}(u, v)\tilde{F}_{jk}(v, w) &= F_{il}(u, s)F_{lj}(s, v)F_{jl}(v, s)F_{lk}(s, w) \\ &= F_{il}(u, s)F_{lk}(s, w) = \tilde{F}_{ik}(u, w), \\ \tilde{F}_{ij}(u, v)\tilde{F}_{ji}(u, v) &= F_{il}(u, s)F_{lj}(s, v)F_{jl}(v, s)F_{li}(s, u) \\ &= F_{il}(u, s)F_{li}(s, u) = 1. \end{aligned}$$

□

5. Proof of the main results.

PROOF OF THEOREM 1. (i) Let Π: $\mathscr{L} \to S$ be a regular, local selection rule, and let $\mathbf{v}^0 = (v_1^0, \ldots, v_n^0)^T = \mathbf{v}^0(\Pi)$. By Lemma 2, for every $\mathbf{v} \neq \mathbf{v}^0$ there exists

a unique $L = L(\mathbf{v}) \in \mathscr{M}$ such that $\Pi(L) = \mathbf{v}$; moreover, $L(\mathbf{v})$ depends continuously on $\mathbf{v}$. Write $L = L(\mathbf{v})$ as

$$L = \begin{cases} \{\mathbf{w}\colon \mathbf{a}^T\mathbf{w} = \mathbf{a}^T\mathbf{v}\}, & \text{if } S = R^n \text{ or } R_+^n, \\ \{\mathbf{w}\colon \mathbf{a}^T\mathbf{w} = \mathbf{a}^T\mathbf{v};\ \mathbf{1}^T\mathbf{w} = 1\}, & \text{if } S = \Delta_n. \end{cases} \tag{5.1}$$

Here the vector $\mathbf{a} \in R^n$ is determined up to equivalence (cf. the passage before Lemma 3).

Our first claim is that there exist continuous functions $g_i(v)$, $v \in V$, with $g_i(v_i^0) = 0$, $i = 1, \ldots, n$, such that in the representation (5.1) of $L = L(\mathbf{v})$ (for arbitrary $\mathbf{v} \neq \mathbf{v}^0$)

$$\mathbf{a} \sim (g_1(v_1), \ldots, g_n(v_n))^T. \tag{5.2}$$

For later reference, observe that (5.2) implies by Lemma 3(i) that

$$v_i \leftrightarrow_{ij} v_j \quad \text{if and only if} \quad g_i(v_i) = g_j(v_j), \tag{5.3}$$

providing in the case $S = \Delta_n$ that $v_i + v_j < 1$.

To prove our first claim, we start with the simpler cases $S = R^n$ or R_+^n. By Lemma 3(ii), applied to $J = \{i, j\}$, it follows that $(a_i, a_j)^T$ is determined up to equivalence by v_i and v_j. Thus

$$F_{ij}(v_i, v_j) = \frac{a_i}{a_j} \tag{5.4}$$

is a well-defined continuous function of v_i and v_j whenever $v_j \neq v_j^0$ [which, by Lemma 3(i), is necessary and sufficient for $a_j \neq 0$]. Clearly, the functions (5.4) satisfy the functional equations

$$F_{ij}(v_i, v_j)F_{jk}(v_j, v_k) = F_{ik}(v_i, v_k), \qquad F_{ij}(v_i, v_j)F_{ji}(v_j, v_i) = 1$$

for $v_i \in V \setminus \{v_i^0\}$, $i = 1, \ldots, n$ (recall that $V = R$ or R_+ according as $S = R^n$ or R_+^n). It follows by Lemma 4 that

$$F_{ij}(v_i, v_j) = \frac{g_i(v_i)}{g_j(v_j)} \tag{5.5}$$

if $v_i \neq v_i^0$, $v_j \neq v_j^0$, for suitable functions defined and not equal to 0 on $V \setminus \{v_i^0\}$. Letting $g_i(v_i^0) = 0$, (5.5) also holds for $v_i = v_i^0$ whenever $F_{ij}(v_i, v_j)$ is defined, that is, $v_j \neq v_j^0$.

Comparing (5.4) and (5.5), we obtain (5.2). The functions g_i are continuous because the F_{ij} are such.

Turning to the more difficult case $S = \Delta_n$, apply Lemma 3(ii) with $J = \{i, j, l\}$ to obtain for $\mathbf{a} \in R^n$ in (5.1) that $(a_i, a_j, a_l)^T$ is uniquely determined, up to equivalence, by v_i, v_j, v_l. Hence

$$F_{ijl}(v_i, v_j, v_l) = \frac{a_i - a_l}{a_j - a_l} \tag{5.6}$$

is a well-defined continuous function of (v_i, v_j, v_l) subject to $v_i + v_j + v_l < 1$,

except for $v_j \leftrightarrow_{jl} v_l$ [which is necessary and sufficient for $a_j = a_l$, by Lemma 3(i)].

Clearly, the functions (5.6) satisfy the functional equations

$$F_{ijl}(v_i, v_j, v_l) F_{jkl}(v_j, v_k, v_l) = F_{ikl}(v_i, v_k, v_l) \tag{5.7}$$

if $v_i + v_j + v_k + v_l < 1$, as well as

$$F_{ijl}(v_i, v_j, v_l) F_{jil}(v_j, v_i, v_l) = 1, \tag{5.8}$$

$$F_{ijl}(v_i, v_j, v_l) F_{jli}(v_j, v_l, v_i) F_{lij}(v_l, v_i, v_j) = -1, \tag{5.9}$$

$$F_{ijl}(v_i, v_j, v_l) + F_{ilj}(v_i, v_l, v_j) = 1 \tag{5.10}$$

if $v_i + v_j + v_l < 1$, assuming in each case that all functions are defined.

These functional equations can be solved applying Lemma 4(ii) three times. First, we use (5.7) and (5.8) fixing l and v_l and restricting the domain of the F_{ijl}'s—as functions of v_i and v_j—by excluding $v_i \leftrightarrow_{il} v_l$, that is, $F_{ijl} = 0$. Then the hypotheses of Lemma 4(ii) are easily checked, taking into account for hypotheses (3) and (4) that (for fixed v_l) the relation $v \leftrightarrow_{il} v_l$ never holds if v is sufficiently small; the latter follows from the continuity axiom in Definition 2. Lemma 4 gives

$$F_{ijl}(v_i, v_j, v_l) = \frac{G_{il}(v_i, v_l)}{G_{jl}(v_j, v_l)}, \tag{5.11}$$

for suitable functions G_{il} defined (and nonzero) for $v_i + v_l < 1$, unless $v_i \leftrightarrow_{il} v_l$.

Substituting (5.11) into (5.9), we obtain, after rearranging,

$$\frac{G_{ji}(v_j, v_i)}{G_{ij}(v_i, v_j)} \frac{G_{lj}(v_l, v_j)}{G_{jl}(v_j, v_l)} = -\frac{G_{li}(v_l, v_i)}{G_{il}(v_i, v_l)}.$$

Now we apply Lemma 4(ii) to the functions

$$H_{il}(v_i, v_l) = -\frac{G_{li}(v_l, v_i)}{G_{il}(v_i, v_l)}$$

(defined for $v_i + v_l < 1$, unless $v_i \leftrightarrow_{il} v_l$), yielding

$$H_{il}(v_i, v_l) = \frac{h_i(v_i)}{h_l(v_l)},$$

for suitable functions h_i defined (and nonzero) in $V = (0, 1)$. This means that the functions

$$\tilde{G}_{il}(v_i, v_l) = \frac{G_{il}(v_i, v_l)}{h_l(v_l)}$$

satisfy

$$\tilde{G}_{il}(v_i, v_l) = -\tilde{G}_{li}(v_l, v_i), \tag{5.12}$$

and, by (5.11),

$$F_{ijl}(v_i, v_j, v_l) = \frac{\tilde{G}_{il}(v_i, v_l)}{\tilde{G}_{jl}(v_j, v_l)}. \tag{5.13}$$

At this point we remove the temporary exclusion of $v_i \leftrightarrow_{il} v_l$ from the definition of the domain of F_{ijl}; defining $\tilde{G}_{il}(v_i, v_l) = 0$ if $v_i \leftrightarrow_{il} v_l$, (5.13) will always hold whenever F_{ijl} in (5.6) is defined.

Finally, substituting (5.13) into (5.10) gives, after rearranging, using also (5.12),

$$\tilde{G}_{il}(v_i, v_l) + \tilde{G}_{lj}(v_l, v_j) = \tilde{G}_{ij}(v_i, v_j), \tag{5.14}$$

whenever $v_i + v_j + v_l < 1$. More exactly, the given derivation of (5.14) is valid unless $v_j \leftrightarrow_{jl} v_l$ and a similar derivation from (5.10), with the roles of i and j interchanged, is valid unless $v_i \leftrightarrow_{il} v_l$; if both $v_i \leftrightarrow_{il} v_l$ and $v_j \leftrightarrow_{jl} v_l$, then also $v_i \leftrightarrow_{ij} v_j$ and (5.14) holds trivially.

(5.14) and (5.12) mean that Lemma 4(ii) is applicable to the functions $F_{ij} = \exp \tilde{G}_{ij}$. It follows that

$$\tilde{G}_{ij}(v_i, v_j) = g_i(v_i) - g_j(v_j), \tag{5.15}$$

for suitable functions g_i defined on $V = (0, 1)$.

Since $\tilde{G}_{ij}(v_i, v_j) = 0$ if $v_i \leftrightarrow_{ij} v_j$ and thus, in particular, if $v_i = v_i^0$, $v_j = v_j^0$, here $g_i(v_i^0)$ is independent of i, and we may assume that $g_i(v_i^0) = 0$, $i = 1, \ldots, n$.

Substituting (5.15) into (5.13), we obtain

$$F_{ijl}(v_i, v_j, v_l) = \frac{g_i(v_i) - g_l(v_l)}{g_j(v_j) - g_l(v_l)}, \tag{5.16}$$

whenever the left-hand side is defined. As the functions F_{ijl} are continuous, so are the g_i's, too.

Comparing (5.16) with (5.6), we obtain (5.2).

Having established our first claim, we define

$$f_i(v) = \int_{v_i^0}^{v} g_i(t)\, dt, \qquad F(\mathbf{v}) = \sum_{i=1}^{n} f_i(v_i). \tag{5.17}$$

We will show that $f_1, \ldots, f_n$ form a standard n-tuple [notice that property (i) in Definition 8 obviously holds] and that $F(\mathbf{v})$ generates II.

With (5.17), (5.2) becomes $\mathbf{a} \sim \operatorname{grad} F(\mathbf{v})$. As the vector $\mathbf{a}$ in (5.1) is determined up to equivalence only, it follows that (for arbitrary $\mathbf{v} \neq \mathbf{v}^0$) $L(\mathbf{v})$ is the set of all $\mathbf{w} \in S$ satisfying

$$(\operatorname{grad} F(\mathbf{v}))^T(\mathbf{w} - \mathbf{v}) = 0. \tag{5.18}$$

Observe next that for any $L \in \mathscr{L}$ we have by the consistency axiom and Lemma 1 that $\Pi(L) = \mathbf{v}$ if and only if $\mathbf{v} \in L \subset L(\mathbf{v})$, providing $\mathbf{v} \neq \mathbf{v}^0$, whereas if $\mathbf{v}^0 \in L$, then always $\Pi(L) = \mathbf{v}^0$. Thus we may assert without any restriction

on $\mathbf{v}$ and for arbitrary $L \in \mathscr{L}$, that $\Pi(L) = \mathbf{v}$ if and only if $\mathbf{v} \in L$ and (5.18) holds for every $\mathbf{w} \in L$ (the latter being automatically fulfilled if $\mathbf{v} = \mathbf{v}^0$).

For any distinct $\mathbf{v}$ and $\mathbf{w}$ in S satisfying (5.18), the result in the last paragraph applied to the line L' through $\mathbf{v}$ and $\mathbf{w}$ gives that

$$(\operatorname{grad} F(\mathbf{w}))^T(\mathbf{w} - \mathbf{v}) \neq 0;$$

here we used the fact that the difference of any two elements of L' is a scalar multiple of $\mathbf{w} - \mathbf{v}$.

By continuity, this nonzero inner product must be of constant sign for $\mathbf{w} \neq \mathbf{v}$ satisfying (5.18), when $\mathbf{v} \in S$ is fixed. Further, again by continuity, this sign cannot actually depend on $\mathbf{v}$, either. Without restricting generality, we may assume that this constant sign is positive. Indeed, the functions g_i in (5.17) may be multiplied by (-1) if necessary, without changing their properties asserted in our first claim.

We have obtained that (5.18) with $\mathbf{w} \neq \mathbf{v}$ always implies

$$(\operatorname{grad} F(\mathbf{w}))^T(\mathbf{w} - \mathbf{v}) > 0. \tag{5.19}$$

It follows that for any line L' in S, $F(\mathbf{w})$ is strictly increasing as $\mathbf{w}$ moves away from $\mathbf{v} = \Pi(L')$ in either direction. This immediately gives that F has property (iii) in Definition 8.

Further, for any $L \in \mathscr{L}$, the fact that $F(\mathbf{w})$ is strictly increasing as $\mathbf{w}$ moves away from $\mathbf{v} = \Pi(L)$ on any line $L' \subset L$ proves that $\Pi(L)$ is the unique point where F is minimized over L. Thus Π is generated by F.

To prove that (ii) in Definition 8 also holds, notice first that as $\mathbf{v} = \mathbf{v}^0$ trivially satisfies (5.18) for every $\mathbf{w} \in S$, (5.19) gives

$$(\operatorname{grad} F(\mathbf{w}))^T(\mathbf{w} - \mathbf{v}^0) > 0 \quad \text{for all } \mathbf{w} \neq \mathbf{v}^0 \text{ in } S. \tag{5.20}$$

If $S = R_+^n$ or Δ_n, let $\hat{\mathbf{v}} = (\hat{v}_1, \ldots, \hat{v}_n)^T$ be a boundary point of S such that $\hat{v}_i = 0$ and $\hat{v}_j \neq 0$ if $j \neq i$, further, $f_j'(\hat{v}_j) \neq f_l'(\hat{v}_l)$ for some j and l different from i. (5.20) implies that

$$\liminf_{\mathbf{w} \to \hat{\mathbf{v}}} (\operatorname{grad} F(\mathbf{w}))^T(\hat{\mathbf{v}} - \mathbf{v}^0) \geq 0;$$

hence $f_i'(v)$ is bounded from above as $v \to 0$. Thus, if the assertion $\lim_{v \to 0} f_i'(v) = -\infty$ were false, f_i' would have a finite limit on a suitable sequence of positive numbers approaching 0. Then there would exist a sequence $\mathbf{v}_k \to \hat{\mathbf{v}}$ (with $\mathbf{v}_k \in S$) such that grad $F(\mathbf{v}_k) \to \mathbf{a}$, say, where $\mathbf{a}$ is not equivalent to $\mathbf{0}$. This provides the desired contradiction because the sequence of subspaces

$$L_k = \left\{\mathbf{w}\colon (\operatorname{grad} F(\mathbf{v}_k))^T(\mathbf{w} - \mathbf{v}_k) = 0, \mathbf{w} \in S\right\} \in \mathscr{M}$$

converges to

$$L = \{\mathbf{w}\colon \mathbf{a}^T(\mathbf{w} - \hat{\mathbf{v}}) = 0, \mathbf{w} \in S\} \in \mathscr{M},$$

and the sequence $\Pi(L_k) = \mathbf{v}_k$ does not converge to $\Pi(L)$, for it goes to $\hat{\mathbf{v}} \notin S$.

This completes the proof of the assertions made in the passage containing (5.17).

Finally, let $f_1, \dots, f_n$ be any standard n-tuple with 0 at $\mathbf{v}^0 = (v_1^0, \dots, v_n^0)$, and set $F(\mathbf{v}) = \sum_{i=1}^n f_i(v_i)$. Notice first that in the cases $S = R^n$ or R_+^n, property (iii) in Definition 8 implies that $f_i(v)$ is strictly increasing (decreasing) for $v > v_i^0$ ($v < v_i^0$) and also that $f_i(v) \to \infty$ as $v \to \infty$. To verify the latter, suppose indirectly that $f_i(v) \to c < \infty$ as $v \to \infty$, choose $v_i > v_i^0$, $v_j > v_j^0$ with $f_i(v_i) + f_j(v_j) = c$, and apply (3.2) to $\mathbf{v} = (v_1, \dots, v_n)^T$ with $v_l = v_l^0$ for $l \neq i, j$ and $\mathbf{v}' = (v_1', \dots, v_n')^T$ with $v_l' = v_l^0$ for $l \neq i$. It follows that

$$f_i(\alpha v_i + (1-\alpha)v_i') + f_j\big(\alpha v_j + (1-\alpha)v_j^0\big) < f_i(v_i) + f_j(v_j), \tag{5.21}$$

for every $v_i' > v_i^0$ and $0 < \alpha < 1$. Letting first $v_i' \to \infty$ and then $\alpha \to 0$, (5.21) results in the contradiction $c \leq f_i(v_i)$. One verifies in the same way that in the case $S = R^n$, $f_i(v) \to \infty$ also as $v \to -\infty$. If $S = R_+^n$ or Δ_n, $F(\mathbf{v})$ has a limit (finite or $+\infty$) as $\mathbf{v}$ converges to a boundary point of S, because property (ii) in Definition 8 implies the monotonicity of each f_i near 0. It follows that F, extended by continuity to the closure of S if $S = R_+^n$ or Δ_n, attains its minimum on any $L \in \mathscr{L}$, more exactly, if $S = R_+^n$ or Δ_n, on the closure of L. But property (ii) rules out the minimum being attained on the boundary. By property (iii), the point where F attains its minimum over L must be unique; hence F does generate a selection rule Π. Property (i) implies that this Π is regular and it is obviously local.

(ii) Let us be given a regular, local projection rule with component projection rules $\Pi(\cdot|\mathbf{u})$, $\mathbf{u} \in S$. For arbitrary $\mathbf{v} \neq \mathbf{u}$, let $L(\mathbf{v}|\mathbf{u})$ denote the unique $L \in \mathscr{M}$ with $\Pi(L|\mathbf{u}) = \mathbf{v}$. We claim that the following modified version of our first claim in the proof of part (i) is valid:

There exist functions $g_i(v|u)$, $u, v \in V$, continuous in v and vanishing for $v = u$, $i = 1, \dots, n$, such that in the representation (5.1) of $L = L(\mathbf{v}|\mathbf{u})$ we have

$$\mathbf{a} \sim \big(g_1(v_1|u_1), \dots, g_n(v_n|u_n)\big)^T. \tag{5.22}$$

In the proof of part (i), the function F generating the selection rule Π was constructed from functions g_i with the property (5.2). If (5.22) is established, the functions $g_i(v|u_i)$ there can be taken as functions $g_i(v)$ corresponding to $\Pi = \Pi(\cdot|\mathbf{u})$. Then [cf. (5.17)] $\Pi(\cdot|\mathbf{u})$ will be generated by

$$F(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^n f_i(v_i|u_i), \qquad f_i(v_i|u_i) = \int_{u_i}^{v_i} g_i(v|u_i)\, dv \tag{5.23}$$

as a function of $\mathbf{v}$. This means, by definition, that the given projection rule is generated by $F(\mathbf{v}|\mathbf{u})$.

Thus it suffices to prove the claim about (5.22). This can be done along the lines of part (i); hence we only sketch the proof, for the hard case $S = \Delta_n$.

We need an obvious modification of Lemma 3, namely that for L as in (4.1) with $\Pi(L|\mathbf{u}) = \mathbf{v} \neq \mathbf{u}$, (a) $(v_i|u_i) \leftrightarrow_{ij} (v_j|u_j)$ iff $a_i = a_j$ and (b) the vector $\mathbf{a}_J$ is determined by $\mathbf{u}_J$ and $\mathbf{v}_J$ up to equivalence.

Applying this to $J = \{i, j, l\}$, it follows that the functions

$$F_{ijl}(v_i, v_j, v_l | u_i, u_j, u_l) = \frac{a_i - a_l}{a_j - a_l} \tag{5.24}$$

are well defined and continuous in v_i, v_j, v_l subject to $u_i + u_j + u_l < 1$, $v_i + v_j + v_l < 1$, except when $(v_j|u_j) \leftrightarrow_{jl} (v_l|u_l)$.

The functions (5.24) satisfy functional equations similar to (5.7)–(5.10), each variable v_i in the latter being replaced by a pair of variables v_i, u_i, where with each constraint on sums of variables v_i a similar constraint on sums of variables u_i is imposed. This system of functional equations can be solved similarly to (5.7)–(5.10), again applying Lemma 4(ii) three times. It is convenient that the variables of the functions in Lemma 4 were not required to be reals; presently we have to let them stand for pairs of real numbers v, u. Finally, we arrive at a representation of the functions (5.24) analogous to (5.16), namely,

$$F_{ijl}(v_i, v_j, v_l | u_i, u_j, u_l) = \frac{g_i(v_i|u_i) - g_l(v_l|u_l)}{g_j(v_j|u_j) - g_l(v_l|u_l)}. \tag{5.25}$$

Comparing (5.24) and (5.25) proves the desired relation (5.22) and thereby part (ii) of Theorem 1.

(iii) Suppose that $F(\mathbf{v}) = \sum_{i=1}^n f_i(v_i)$ and $\tilde{F}(\mathbf{v}) = \sum_{i=1}^n \tilde{f}_i(v_i)$ generate the same selection rule, where $(f_1, \ldots, f_n)$ and $(\tilde{f}_1, \ldots, \tilde{f}_n)$ are standard n-tuples. Clearly, it suffices to prove $\tilde{f}_i = cf_i$ for the case when $(\tilde{f}_1, \ldots, \tilde{f}_n)$ is arbitrary, with 0 at $\mathbf{v}^0 = (v_1^0, \ldots, v_n^0)^T$, say, and $(f_1, \ldots, f_n)$ is constructed to the (regular, local) selection rule Π generated by F as in the proof of part (i) [cf. (5.17)]. Now, since $\tilde{F}$ generates Π, its minimum on $L = L(\mathbf{v})$ is achieved at the point $\mathbf{v}$. Hence for every $\mathbf{v} \neq \mathbf{v}^0$ we have $(\text{grad } \tilde{F}(\mathbf{v}))^T(\mathbf{w} - \mathbf{v}) = 0$ for all $\mathbf{w} \in L(\mathbf{v})$. Since $L(\mathbf{v})$ is the set of all $\mathbf{w} \in S$ satisfying (5.18), it follows that for $\mathbf{v} \neq \mathbf{v}^0$,

$$\text{grad } \tilde{F}(\mathbf{v}) = \lambda(\mathbf{v}) \text{grad } F(\mathbf{v}) \quad \text{if } S = R^n \text{ or } R_+^n, \tag{5.26}$$

$$\text{grad } \tilde{F}(\mathbf{v}) = \lambda(\mathbf{v}) \text{grad } F(\mathbf{v}) + \mu(\mathbf{v})\mathbf{1} \quad \text{if } S = \Delta_n; \tag{5.27}$$

the same holds trivially also for $\mathbf{v} = \mathbf{v}^0$, with $\mu(\mathbf{v}^0) = 0$ in the case $S = \Delta_n$.

As the components of the gradient vectors depend only on the corresponding components of $\mathbf{v}$, the scalar functions λ and μ in (5.26) and (5.27) must be constant and, in particular, μ in (5.27) is identically 0. Thus we actually have

$$\text{grad } \tilde{F}(\mathbf{v}) = c \text{ grad } F(\mathbf{v}). \tag{5.28}$$

Since $f_1, \ldots, f_n$ and $\tilde{f}_1, \ldots, \tilde{f}_n$ have the property (i) in Definition 8, (5.28) implies that $\tilde{f}_i = cf_i$, as claimed. The proof of assertion (iii) for projection rules is similar. □

PROOF OF THEOREM 2. By the proof of Theorem 1(ii), any regular and local projection rule is generated by a function as in (5.23), where $g_i(v|u)$ is a continuous function of v that vanishes at $v = u$, $i = 1, \ldots, n$; moreover, for any $\mathbf{u} \neq \mathbf{v}$ the subspace L defined by (5.1) with $a_i = g_i(v_i|u_i)$ has the property

that $\Pi(L|\mathbf{u}) = \mathbf{v}$. By the modification of Lemma 3(i) used in that proof, it follows that

$$(5.29) \qquad (v_i|u_i) \leftrightarrow_{ij} (v_j|u_j) \quad \text{if and only if} \quad g_i(v_i|u_i) = g_j(v_j|u_j),$$

provided in the case $S = \Delta_n$ that $u_i + u_j < 1$, $v_i + v_j < 1$.

Now, if $S = R^n$ or R_+^n and the given projection rule is semisymmetric, that is, $(v|u) \leftrightarrow_{ij} (v|u)$ for every u and v in V and every $i, j \in \{1, \dots, n\}$, (5.29) implies that the functions g_i do not depend on i; hence, by (5.23), neither do the functions f_i.

Further, if $S = R_+^n$ or Δ_n and the given projection rule is statistical, that is, $(v|u) \leftrightarrow_{ij} (v'|u')$ if and only if $v/u = v'/u'$, (5.29) implies that

$$(5.30) \qquad g_i(v_i|u_i) = g_j(v_j|u_j) \quad \text{if } \frac{v_i}{u_i} = \frac{v_j}{u_j},$$

provided in the case $S = \Delta_n$ that $u_i + u_j < 1$, $v_i + v_j < 1$. Actually, the last constraint can be dispensed with, because for any given u_i, u_j, v_i, v_j [in $V = (0, 1)$] there exist u_k, v_k such that $u_i + u_k < 1$, $v_i + v_k < 1$, $u_j + u_k < 1$, $v_j + v_k < 1$ and $v_i/u_i = v_k/u_k$. (5.30) means that $g_i(v|u)$ is a one-to-one function of v/u, not depending on i, that is,

$$(5.31) \qquad g_i(v|u) = g\left(\frac{v}{u}\right).$$

The continuity of g_i as a function of v and the one-to-one property of g implies that $g(t)$ is a continuous, strictly monotonic function of t, and $g(1) = g_i(u|u) = 0$. Substituting (5.31) into (5.26) gives

$$(5.32) \qquad f_i(v_i|u_i) = \int_{u_i}^{v_i} g\left(\frac{v}{u_i}\right) dv = u_i f\left(\frac{v_i}{u_i}\right), \qquad f(t) = \int_1^t g(s)\, ds.$$

This completes the proof of Theorem 2. □

PROOF OF THEOREM 3. (i) First, we show that if a regular selection rule is subspace-transitive, then for distinct elements $\mathbf{u}, \mathbf{v}, \mathbf{w}$ of S,

$$(5.33) \qquad L(\mathbf{v}|\mathbf{u}) \cap L(\mathbf{w}|\mathbf{v}) \subset L(\mathbf{w}|\mathbf{u}) \quad \text{if } \mathbf{w} \in L(\mathbf{v}|\mathbf{u}).$$

Here, as in the proof of Theorem 1(ii), $L(\mathbf{v}|\mathbf{u})$ denotes the unique $L \in \mathscr{M}$ with $\Pi(L|\mathbf{u}) = \mathbf{v}$.

Indeed, let $L' = L(\mathbf{v}|\mathbf{u}) \cap L(\mathbf{w}|\mathbf{v})$. Then $\Pi(L'|\mathbf{v}) = \mathbf{w}$ by the consistency axiom, and the transitivity postulate applied to $L' \subset L = L(\mathbf{v}|\mathbf{u})$ gives

$$\Pi(L'|\mathbf{u}) = \Pi(L'|\mathbf{v}) = \mathbf{w}.$$

But $\Pi(L'|\mathbf{u}) = \mathbf{w}$ implies that $L' \subset L(\mathbf{w}|\mathbf{u})$, by the corollary of Lemma 1, proving (5.33).

Now let us be given a regular, local projection rule, generated by $F(\mathbf{v}|\mathbf{u})$ as in (5.23). In particular,

$$g_i(v|u) = \frac{\partial}{\partial v} f_i(v|u) \tag{5.34}$$

is a continuous function of v, vanishing at $v = u$. Then $L(\mathbf{v}|\mathbf{u})$ consists of those $\mathbf{w} \in S$ that satisfy

$$\sum_{i=1}^{n} g_i(v_i|u_i)(w_i - v_i) = 0. \tag{5.35}$$

Hence, if the given projection rule is subspace-transitive, it follows from (5.33) that for any distinct $\mathbf{u}, \mathbf{v}, \mathbf{w}$ in S satisfying (5.35), there exist scalars α, β, γ, possibly depending on $\mathbf{u}, \mathbf{v}, \mathbf{w}$, with $\gamma = 0$ unless $S = \Delta_n$, such that

$$\alpha g_i(v_i|u_i) + \beta g_i(w_i|v_i) + \gamma = g_i(w_i|u_i), \qquad i = 1, \ldots, n. \tag{5.36}$$

We claim that actually

$$g_i(v|u) + g_i(w|v) = g_i(w|u), \tag{5.37}$$

for every u, v, w in V and $i = 1, \ldots, n$. Clearly, it suffices to prove this for $i = 1$.

Consider first the simpler cases $S = R^n$ or R_+^n. Then for any u, v, w in V, there exist $\mathbf{u}, \mathbf{v}, \mathbf{w}$ in S satisfying (5.35) such that $u_1 = u$, $v_1 = v$, $w_1 = w$ and, in addition,

$$u_2 = v_2 \neq w_2, \qquad u_3 \neq v_3 = w_3. \tag{5.38}$$

With these $\mathbf{u}, \mathbf{v}, \mathbf{w}$, (5.38) implies that in (5.36), where now $\gamma = 0$, we have $\alpha = \beta = 1$ [using that, by Lemma 3, $g_i(v|u) \neq 0$ if $v \neq u$]. This proves (5.37) for $i = 1$.

If $S = \Delta_n$ then $v \neq u$ does not necessarily imply $g_i(v|u) \neq 0$. It follows, however, from (iii) in Definition 8, that for at most one index i can $u < 1/2$ and an interval $I \subset (0, 1/2)$ be found such that $f_i(v|u)$ is constant for $v \in I$. This implies (assuming, without any loss of generality, that the exceptional i, if any, is different from 2 and 3) that for any $\delta < 1/2$, the numbers ε satisfying

$$g_2(\varepsilon|\delta) \neq 0, \quad g_3(\varepsilon|\delta) \neq 0 \tag{5.39}$$

are dense in the interval $(0, 1/2)$. Using this, it is easy to see that for any fixed u, v and w sufficiently close to v, there exist $\mathbf{u}, \mathbf{v}, \mathbf{w}$ in $S = \Delta_n$ satisfying (5.35), such that $u_1 = u$, $v_1 = v$, $w_1 = w$ and, with some ε and δ satisfying (5.39),

$$u_2 = v_2 = u_3 = \delta, \qquad w_2 = v_3 = w_3 = \varepsilon \tag{5.40}$$

and

$$u_4 = v_4 = w_4. \tag{5.41}$$

With these $\mathbf{u}, \mathbf{v}, \mathbf{w}$, (5.41) implies that in (5.36) we have $\gamma = 0$, then (5.40) and

(5.39) imply that $\alpha = \beta = 1$. This proves (5.37) (for $i = 1$) if w is sufficiently close to v.

Observe next that given any u and v in $V = (0, 1)$, the numbers $w \in V$ satisfying (5.37) for $i = 1$ form an open set. Indeed, by the last paragraph, for w' sufficiently close to w we have

$$g_1(w'|w) + g_1(w|v) = g_1(w'|v), \qquad g_1(w'|w) + g_1(w|u) = g_1(w'|u);$$

it follows that if w satisfies (5.37) for $i = 1$, then so does w'.

Since $g_1(w|u)$ and $g_1(w|v)$ are continuous functions of w, the (nonvoid) set of those $w \in V$ that satisfy (5.37) for $i = 1$ can be open only if it equals the whole $V = (0, 1)$. This completes the proof of (5.37) in the case $S = \Delta_n$.

The functional equations (5.37) imply that the functions g_i can be represented as

$$g_i(v|u) = \psi_i(v) - \psi_i(u), \tag{5.42}$$

where it may be assumed that $\psi_i(s) = 0$, $i = 1, \dots, n$, for some $s \in V$ [set $\psi_i(v) = g_i(s|v)$, say]. Since $g_i(v|u)$ is a continuous function of v, ψ_i must be continuous, too. Writing

$$\varphi_i(v) = \int_s^v \psi_i(t)\, dt, \tag{5.43}$$

the functions $\varphi_1, \dots, \varphi_n$ satisfy (i) in Definition 8, with $\mathbf{v}^0 = s \cdot \mathbf{1}$, and (in the cases $S = R_+^n$ or Δ_n) the validity of (ii) in Definition 8 for the functions φ_i follows from that for $g_i(\cdot|u)$, by (5.42). Property (iii) for $\varphi_1, \dots, \varphi_n$ will, of course, follow from the strict convexity of $\Phi(\mathbf{v}) = \sum_{i=1}^n \varphi_i(v_i)$ that we are going to verify immediately.

From (5.34), (5.42) and (5.43) we get

$$f_i(v|u) = \int_u^v g_i(t|u)\, dt = \varphi_i(v) - \varphi_i(u) - \psi_i(u)(v - u). \tag{5.44}$$

This proves that $F(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^n f_i(v_i|u_i)$ has the claimed form

$$F(\mathbf{v}|\mathbf{u}) = \Phi(\mathbf{v}) - \Phi(\mathbf{u}) - (\operatorname{grad} \Phi(\mathbf{u}))^T(\mathbf{v} - \mathbf{u}).$$

Since $F(\mathbf{v}|\mathbf{u}) \geq 0$, with equality if and only if $\mathbf{v} = \mathbf{u}$, this result also proves the strict convexity of Φ on S.

(ii) Let $\varphi_1, \dots, \varphi_n$ be a standard n-tuple such that $\Phi(\mathbf{v}) = \sum_{i=1}^n \varphi_i(v_i)$ is strictly convex on S, and let $f_i(v|u)$ be defined by (5.43) where $\psi_i(u) = \varphi_i'(u)$. Then for any fixed $\mathbf{u} \in S$, the functions $f_i(\cdot|u_i)$ form a standard n-tuple with 0 at $\mathbf{u}$; hence by Theorem 1(ii), $F(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^n f_i(v_i|u_i)$ generates a regular, local projection rule. To prove that the latter is parallel-transitive, suppose that $\mathbf{v} = \Pi(L|\mathbf{u})$, $\mathbf{w} = \Pi(L'|\mathbf{v})$, where L and L' are "parallel" subspaces. Then grad $F(\mathbf{v}|\mathbf{u})$ and grad $F(\mathbf{w}|\mathbf{v})$ are both orthogonal to these subspaces (where the gradient refers to the first variable) and hence so is

$$\operatorname{grad} F(\mathbf{w}|\mathbf{u}) = \operatorname{grad} F(\mathbf{v}|\mathbf{u}) + \operatorname{grad} F(\mathbf{w}|\mathbf{v}).$$

This means that $\Pi(L'|\mathbf{u}) = \mathbf{w}$, as claimed.

To prove the corollary, recall that by Theorem 2, every statistical projection rule is generated by an f-divergence

$$F(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^{n} v_i f\left(\frac{v_i}{u_i}\right),$$

where $f(t)$ is a continuously differentiable function with $f(1) = f'(1) = 0$. Then the functions $g_i(v|u)$ in (5.34) do not depend on i and are equal to $g(v/u)$, where $g(t) = f'(t)$. If this projection rule is transitive, we must have

$$g\left(\frac{v}{u}\right) = \psi(v) \dot{-} \psi(u)$$

[cf. (5.42)] for some continuous function ψ. This implies that g satisfies the functional equation

$$g(ts) = g(t) + g(s), \qquad t, s \in R_+, \tag{5.45}$$

whose only continuous solutions are $g(t) = c \log t$ [cf. Aczél (1966), Section 2.1.2]. Hence $f(t) = \int_1^t g(t)\,dt = c(t \log t - t)$ and this means that $F(\mathbf{v}|\mathbf{u}) = cI(\mathbf{v}\|\mathbf{u})$. □

PROOF OF THEOREM 4. (i) First we show that a (regular, local) projection rule with basic set $S = R^n$, generated by

$$F(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^{n} f_i(v_i|u_i) \tag{5.46}$$

as in Theorem 1(ii), is translation-invariant if and only if

$$f_i(v + \mu|u + \mu) = c(\mu) f_i(v|u), \tag{5.47}$$

for every u, v and μ in R, where $c(\mu)$ is a suitable positive-valued function. Indeed, since $\mathbf{v}^* = \Pi(L + \mu\mathbf{1}|\mathbf{u} + \mu\mathbf{1})$ minimizes $F(\mathbf{v}|\mathbf{u} + \mu\mathbf{1})$ subject to $\mathbf{v} \in L + \mu\mathbf{1}$, $\mathbf{v}^* - \mu\mathbf{1}$ minimizes $F_\mu(\mathbf{v}|\mathbf{u}) = F(\mathbf{v} + \mu\mathbf{1}|\mathbf{u} + \mu\mathbf{1})$ subject to $\mathbf{v} \in L$. Hence the given projection rule is translation-invariant, that is, $\Pi(L|\mathbf{u}) = \mathbf{v}^* - \mu\mathbf{1}$, if and only if this projection rule is also generated by $F_\mu(\mathbf{v}|\mathbf{u})$. The latter is, by Theorem 1(iii), equivalent to (5.47), as claimed.

By Theorem 3, if (5.46) generates a transitive projection rule, then

$$f_i(v|u) = \varphi_i(v) - \varphi_i(u) - \varphi_i'(u)(v - u), \tag{5.48}$$

where $(\varphi_1, \ldots, \varphi_n)$ is a standard n-tuple. By the paragraph containing (5.43), we may assume without any loss of generality that this standard n-tuple has 0 at $\mathbf{v}^0 = \mathbf{0}$. Our next goal is to determine what functions $f_i(v|u)$ of form (5.48) satisfy (5.47).

Observe that (5.47) implies $c(\mu_1 + \mu_2) = c(\mu_1)c(\mu_2)$ and that (5.47) and (5.48) imply the continuity of $c(\mu)$. It follows that

$$c(\mu) = e^{\beta\mu} \quad \text{for some } \beta \in R \tag{5.49}$$

[cf. Aczél (1966), Section 2.1.2].

From (5.47) (with $\mu = -u$) and (5.49), we obtain

$$f_i(v|u) = e^{\beta u} f_i(v - u|0); \tag{5.50}$$

this and (5.48)—where, by assumption, $\varphi_i(0) = \varphi_i'(0) = 0$— result in

$$f_i(v|u) = e^{\beta u}\varphi_i(v - u). \tag{5.51}$$

Comparing (5.48) and (5.51) and differentiating by v, it follows that

$$\varphi_i'(v) - \varphi_i'(u) = e^{\beta u}\varphi_i'(v - u). \tag{5.52}$$

This means, substituting $v = u + t$, that $\psi = \varphi_i'$ satisfies the functional equation

$$\psi(u + t) = \psi(u) + e^{\beta u}\psi(t). \tag{5.53}$$

This functional equation is solved easily. First,

$$\psi(t) = at \quad \text{if } \beta = 0 \tag{5.54}$$

[cf. Aczél (1966), Section 2.1.2]. For $\beta \neq 0$, observe that (5.53) implies by symmetry

$$\psi(u) + e^{\beta u}\psi(t) = \psi(t) + e^{\beta t}\psi(u);$$

thus

$$\frac{\psi(t)}{\psi(u)} = \frac{e^{\beta t} - 1}{e^{\beta u} - 1}.$$

This gives

$$\psi(t) = a(e^{\beta t} - 1) \quad \text{if } \beta \neq 0. \tag{5.55}$$

Thus φ_i' must be either of form (5.54) or of form (5.55), where the constant factor a (but not β) may depend on i. Clearly, the positivity or negativity of these factors, according as $\beta \geq 0$ or $\beta < 0$, is necessary and sufficient for getting a standard n-tuple $\varphi_1, \ldots, \varphi_n$.

Finally, just as (5.47) was necessary and sufficient for translation invariance, one sees that the projection rule generated by (5.46) is scale-invariant if and only if

$$f_i(\lambda v|\lambda u) = c(\lambda) f_i(v|u), \tag{5.56}$$

for every $\lambda > 0$. Now, $f_i(v|u)$ defined by (5.48) with φ_i' of form (5.55) does not satisfy (5.56), whereas with φ_i' of form (5.54) it does. In the latter case (5.46) becomes

$$F(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^{n} a_i(v_i - u_i)^2, \qquad a_i > 0, i = 1, \ldots, n, \tag{5.57}$$

that is, only the weighted least squares projection rules are transitive as well as location- and scale-invariant.

By Theorem 2(i), the additional postulate of semisymmetry implies that all coefficients a_i in (5.57) must be equal.

(ii) Suppose now that $S = R_+^n$ and a projection rule generated by a function as in (5.46) is transitive and scale-invariant. Then (5.48) and (5.56) hold; in the former we now suppose that the standard n-tuple $(\varphi_1, \ldots, \varphi_n)$ has 0 at $\mathbf{v}^0 = \mathbf{1}$.

As (5.56) implies that $c(\lambda_1\lambda_2) = c(\lambda_1)c(\lambda_2)$ and (5.48) and (5.56) imply the continuity of $c(\lambda)$, we have $c(\lambda) = \lambda^\alpha$ for some $\alpha \in R$ [Aczél (1966), Section 2.1.2]. Hence from (5.56) (with $\lambda = 1/u$) and (5.48) [where now $\varphi_i(1) = \varphi_i'(1) = 0$], we obtain

$$f_i(v|u) = u^\alpha f_i\left(\frac{v}{u}\middle|1\right) = u^\alpha \varphi_i\left(\frac{v}{u}\right). \tag{5.58}$$

Comparing (5.48) and (5.58) and differentiating by v, it follows that

$$\varphi_i'(v) - \varphi_i'(u) = u^{\alpha-1}\varphi_i'\left(\frac{v}{u}\right)$$

or, substituting $v = tu$,

$$\varphi_i'(tu) = \varphi_i'(u) + u^{\alpha-1}\varphi_i'(t). \tag{5.59}$$

Since (5.59) means that $\psi(t) = \varphi_i'(e^t)$ satisfies the functional equation (5.53) (with $\beta = \alpha - 1$), we obtain from (5.54) and (5.55)

$$\varphi_i'(t) = \begin{cases} a_i \log t, & \text{if } \alpha = 1, \\ a_i(t^{\alpha-1} - 1), & \text{if } \alpha \neq 1. \end{cases} \tag{5.60}$$

Clearly, (5.60) with $\varphi_i(1) = 0$ defines a standard n-tuple for $S = R_+^n$ iff $\alpha \leq 1$ and, in addition, $a_i > 0$ in the case $\alpha = 1$ or $a_i < 0$ in the case $\alpha < 1$. With this choice of φ_i, (5.48) becomes $f_i(v|u) = |a_i|h_\alpha(v|u)$, with h_α defined by (3.7). On the other hand, (5.46) with these functions f_i does generate a transitive and scale-invariant projection rule. If this projection rule is also semisymmetric, it follows by Theorem 2(i) that it is generated by (3.8). The proof is complete. □

PROOF OF THEOREM 5. Suppose that $S = R^{mn}$, R_+^{mn} or Δ_{mn}, the elements of S being represented as $\mathbf{v} = \{v_{ij}\}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$. Let Π be a regular, local selection rule with basic set S.

By Theorem 1, Π is generated by a function

$$F(\mathbf{v}) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(v_{ij}), \tag{5.61}$$

where the functions f_{ij} form a standard mn-tuple with 0 at $\mathbf{v}^0 = \{v_{ij}^0\} = \mathbf{v}^0(\Pi)$. In particular, the functions $g_{ij}(t) = (d/dt)f_{ij}(t)$ are continuous and

$$f_{ij}(v_{ij}^0) = g_{ij}(v_{ij}^0) = 0. \tag{5.62}$$

Consider the subspaces $L_{\mathbf{v}}$ defined by (2.13), that is, $L_{\mathbf{v}} = \{\mathbf{w}\colon \overline{\mathbf{w}} = \overline{\mathbf{v}};\ \overline{\overline{\mathbf{w}}} = \overline{\overline{\mathbf{v}}}\}$. Then if $\Pi(L_{\mathbf{v}}) = \mathbf{v}$ for some $\mathbf{v} \in S$, that is, if the minimum of (5.61) on $L_{\mathbf{v}}$ is attained at $\mathbf{v}$, we have

$$g_{ij}(v_{ij}) = \lambda_i + \mu_j, \qquad i = 1, \ldots, m,\ j = 1, \ldots, n, \tag{5.63}$$

for suitable "Lagrange multipliers" λ_i, μ_j. It follows from (5.63) that

$$g_{ij}(v_{ij}) + g_{kl}(v_{kl}) = g_{il}(v_{il}) + g_{kj}(v_{kj}), \tag{5.64}$$

for every i, j, k, l.

Now we proceed to prove part (i) of Theorem 5. If Π is sum-consistent, then by Definition 7, $\Pi(L_{\mathbf{v}}) = \mathbf{v}$ always holds if $\mathbf{v}$ is of sum form, $v_{ij} = s_i + t_j$. Thus (5.64) gives the system of functional equations

$$g_{ij}(s_i + t_j) + g_{kl}(s_k + t_l) = g_{il}(s_i + t_l) + g_{kj}(s_k + t_j). \tag{5.65}$$

Observe first that (5.65) implies the differentiability of the (continuous) functions g_{ij}; this can be seen, for example, by integrating both sides of (5.65) with respect to s_k.

Differentiating both sides of (5.64) by t_j on the one hand and by s_k on the other, we obtain

$$g'_{ij}(s_i + t_j) = g'_{kj}(s_k + t_j) = g'_{kl}(s_k + t_l).$$

This means that g'_{ij} equals the same constant c for every i, j; hence

$$g_{ij}(v) = cv + d_{ij}. \tag{5.66}$$

Recalling (5.62), it follows that

$$g_{ij}(v) = c\big(v - v^0_{ij}\big), \qquad f_{ij}(v) = \frac{c}{2}\big(v - v^0_{ij}\big)^2. \tag{5.67}$$

This proves that if Π is sum-consistent, it must be a least squares selection rule. Recall that, as remarked after Definition 7, $\mathbf{v}^0$ must be of sum form. Conversely, it is easy to see that the least squares selection rules with $\mathbf{v}^0$ of sum form are sum-consistent.

The result just proved easily implies that the only (regular, local) sum-consistent projection rule is the least squares projection rule. In fact, let this projection rule be generated by

$$F(\mathbf{v}|\mathbf{u}) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(v_{ij}|u_{ij}),$$

where the functions $f_{ij}(\cdot|u_{ij})$ form a standard n-tuple with 0 at $\mathbf{u}$ for every fixed $\mathbf{u} = \{u_{ij}\} \in S$. Then (5.67) gives that if $\mathbf{u}$ is of sum form, the terms of this standard mn-tuple must be constant multiples of $(v - u_{ij})^2$. Since for any fixed u there exists $\mathbf{u} = \{u_{ij}\}$ of sum form with $u_{ij} = u$, it follows that $f_{ij}(v|u)$ always equals a constant times $(v - u)^2$.

The proof of part (ii) is similar. If Π generated by (5.61) is product-consistent, then by Definition 8, $\Pi(L_{\mathbf{v}}) = \mathbf{v}$ always holds if $\mathbf{v}$ is of product form, $v_{ij} = s_i t_j$. Thus (5.64) gives

$$g_{ij}(s_i t_j) + g_{kl}(s_k t_l) = g_{il}(s_i t_l) + g_{kj}(s_k t_j). \tag{5.68}$$

Notice that whereas in the case $S = \Delta_{mn}$ the condition $\Sigma v_{ij} = 1$ represents a constraint on the permissible s_i and t_j in (5.68), the equation must certainly be valid—for any fixed i, j, k, l—if $s_i + s_k < 1$, $t_j + t_l < 1$.

The system of functional equations (5.68) can be transformed to (5.65). Namely, if the functions $g_{ij}(v)$ satisfy (5.68), then $\tilde{g}_{ij}(v) = g_{ij}(e^v)$ satisfy (5.65). Hence, from (5.66),

$$g_{ij}(v) = \tilde{g}_{ij}(\log v) = c \log v + d_{ij}. \tag{5.69}$$

Recalling (5.62), it follows that

$$g_{ij}(v) = c \log \frac{v}{v_{ij}^0}, \qquad f_{ij}(v) = c\left[v \log \frac{v}{v_{ij}^0} - v + v_{ij}^0\right]. \tag{5.70}$$

This proves that a product-consistent selection rule must be an I-divergence selection rule (cf. Example 2) with $\mathbf{v}^0$ of product form. Conversely, it is easy to see that these I-divergence selection rules are product-consistent. The assertion that the only product-consistent projection rule is the I-divergence projection rule follows in the same way as did its analog in part (i). The corollary is immediate. □

REFERENCES

ACZÉL, J. (1966). *Lectures on Functional Equations and Their Applications*. Academic, New York.

ACZÉL, J. and DARÓCZY, Z. (1975). *On Measures of Information and Their Characterizations*. Academic, New York.

ALI, S. M. and SILVEY, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* **28** 131–142.

BREGMAN, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comput. Math. and Math. Phys.* **7** 200–217.

VAN CAMPENHOUT, J. M. and COVER, T. M. (1981). Maximum entropy and conditional probability. *IEEE Trans. Inform. Theory* **IT-27** 483–489.

CENSOR, Y. (1983). Finite series-expansion reconstruction methods. *Proceedings of the IEEE* **71** 409–419.

CENSOR, Y. and LENT, A. (1981). An iterative row-action method for interval convex programming. *J. Optim. Theory Appl.* **34** 321–353.

CSISZÁR, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **8** 85–108.

CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318.

CSISZÁR, I. (1984). Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.* **12** 768–793.

CSISZÁR, I. (1985). An extended maximum entropy principle and a Bayesian justification (with discussion). In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 83–89. North-Holland, Amsterdam.

CSISZÁR, I. and TUSNÁDY, G. (1984). Information geometry and alternating minimization procedures. *Statist. Decisions Suppl.* **1** 205–237.

DEMPSTER, A. D., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–37.

DIACONIS, P. and ZABELL, S. L. (1982). Updating subjective probability. *J. Amer. Statist. Assoc.* **77** 831–834.

HERMAN, G. T. and LENT, A. (1976). Iterative reconstruction algorithms. *Computers in Biology and Medicine* **6** 273–294.

ITAKURA, F. and SAITO, S. (1968). Analysis synthesis telephony based on the maximum likelihood method. In *Reports of the Sixth International Congress on Acoustics* (Y. Kohasi, ed.) 17–20. Tokyo, Japan.

JAYNES, E. T. (1982). On the rationale of maximum entropy methods. *Proceedings of the IEEE* **70** 939–952.

JONES, L. K. (1989). Approximation theoretic derivation of logarithmic entropy principles for inverse problems and unique extension of the maximum entropy method to incorporate prior knowledge. *SIAM J. Appl. Math.* **49** 650–661.

JONES, L. K. and BYRNE, C. L. (1990). General entropy criteria for inverse problems, with applications to data compression, pattern classification and cluster analysis. *IEEE Trans. Inform. Theory* **IT-36** 23–30.

JONES, L. and TRUTZER, V. (1989). Computationally feasible high-resolution minimum-distance procedures which extend the maximum-entropy method. *Inverse Problems* **5** 749–766.

KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York.

LIESE, F. and VAJDA, I. (1987). *Convex Statistical Distances*. Teubner, Leipzig.

MILLER, M. I. and SNYDER, D. L. (1987). The role of likelihood and entropy in incomplete-data problems: Applications to estimating point-process intensities and Toeplitz constrained covariances. *Proceedings of the IEEE* **75** 892–907.

PARIS, J. B. and VENCOVSKÁ, A. (1990). A note on the inevitability of maximum entropy. *International Journal of Inexact Reasoning* **4** 183–223.

PEREZ, A. (1984). "Barycenter" of a set of probability measures and its application in statistical decision. *Compstat Lectures* 154–159. Physica, Heidelberg.

SHORE, J. E. and JOHNSON, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum-cross entropy. *IEEE Trans. Inform. Theory* **IT-26** 26–37. [Correction **IT-29** (1983) 942–943.]

SKILLING, J. (1988). The axioms of maximum entropy. In *Maximum Entropy and Bayesian Methods in Science and Engineering* **1** 173–187. Kluwer, Amsterdam.

VARDI, Y., SHEPP, L. A. and KAUFMAN, L. (1985). A statistical model for positron emission tomography (with discussion). *J. Amer. Statist. Assoc.* **80** 8–35.

MATHEMATICAL INSTITUTE OF THE
HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST, P.O.B. 127
H-1364
HUNGARY