MODELS AS INSTRUMENTS, WITH APPLICATIONS TO MOMENT STRUCTURE ANALYSIS

JAN DE LEEUW

ABSTRACT. The paper discusses a very general technique to evaluate the quality of estimates, and of models that these estimates are based on. Methods are based on a simple geometrical argument, and on expansions of the loss functions around the estimate, the target, and the replication. We give both Delta-method and Jackknife computational procedures to estimate the relevant quantities.

1. INTRODUCTION

In De Leeuw [7] several techniques for comparison of estimates and models in multinomial experiments are discussed. These techniques differ from the usual ones, because they do not assume that a particular unsaturated model is true (even approximately). They also are not based on any preferred loss function (such as the log-likelihood) or any preferred estimation principle (such as maximum likelihood). Moreover the methods have computer-intensive versions which do not require evaluating complicated expansion formulas, but use the Jackknife instead. In this paper I continue the work of [7]. I use a geometric framework, in keeping with the recent work on the use of differential geometry in statistics, and with the recent emphasis on minimum distance estimation. See, for instance, the impressive book by Bickel et al. [2], or the recent paper by Lindsay [11]. My approach is much more pedestrian, however, and it uses models (if at all) in a thoroughly instrumentalist way, as devices to improve estimates.

As mentioned above, the results in De Leeuw [7] are limited to multinomial experiments. Moreover, they concentrate on minimum distance estimates, which means that the notion of a restrictive model for the proportions continues to play an important role. In this paper I generalize the more interesting results to a larger class of estimators, based on fairly arbitrary sequences of statistics (which could be proportions, means, covariances, correlations, and so on). These sequences, which are called *data*, converge in probability to a limit value, the *target.* The target is sometimes also called the *truth,* but as you perhaps already have guessed, I do not like that term.

In this paper, I distinguish various types of errors an estimate can have, and I estimate these errors using Delta method and Jackknife techniques. The notion of a distance-type loss function still plays a role. I evaluate the distance between the target and the sequence of estimates. This is called the *overall error.* I also compute the distance between the estimates and a sequence of variables with the same distribution as the target sequence, but independent of the target sequence. This is called the *prediction error.* Finally I compute estimates of the distance between the target and the limit of the estimation sequence, which is the *specification error.*

Written while visiting the National Institute of Statistical Sciences, Research Triangle Park, NC.

Finally, in the computations I need the distance between the target sequence and the estimation sequence, which I call the *deviance.* Just words, of course, but chosen to suggest the various types of errors that are involved.

The paper is closely related to the work of Linhart and Zucchini [12], which has been used by Cudeck and Brown [5] and Browne and Cudeck [3, 4]. A related Bayesian approach, also based on predictive distribution, has been discussed by Gelfand, Dey, and Chang [8]. Related research in covariance structure modeling has been published by McDonald [13], McDonald and Marsh [14], and Bentler [1].

2. Estimation Problem

We start, as in De Leeuw [7], with simple experiments in which we estimate a vector of proportions (this could also be a multidimensional cross table, or a discreticized multivariate distribution). This will be generalized slightly in a later section.

Thus we start by considering an *estimation problem* with the following ingredients.

Definition 2.1. The *data* is sequence of random vectors of *proportions* $\underline{p}_n$, taking values in the unit simplex $\mathbb{S}^m$. The *target* is a fixed element π of $\mathbb{S}^m$.

Of course this is highly suggestive, and to some extent misleading, terminology. Actual data are not a random variable, and certainly not a sequence of random variables. To arrive at this representation we must first embed the actual data in a *framework of replication*, which is modeled by a single random variable, and we then have to embed that random variable in a sequence to get our asymptotics under way. Thus the distance from what we call "data" to what an ordinary person would call "data" is quite large. In the same way calling a particular point in parameter space "the target" must be seen as have some heuristic value, at the most.

We still have to connect the data and the target in some convenient way. The three assumptions below are enough to justify our expansions and expectations, and they also serve to simplify the notation.

Assumption 2.2. For all $n \in \mathbb{N}$ we have $\mathbf{E}(\underline{p}_n) = \pi$.

Assumption 2.3. There is a positive semi-definite $V(\pi)$ such that for all $n \in \mathbb{N}$ we have $n\mathbf{V}(\underline{p}_n) = V(\pi)$. $V(\pi)$ is of rank $m-1$, and only has the constant vectors in its null-space.

Assumption 2.4. For all $j = 1, \cdots, m$ we have $\mathbf{E} \mid \underline{p}_{jn} - \pi_j \mid^6 = \mathcal{O}(n^{-3})$

By Liaponoff's Inequality on absolute moments, Assumption 2.4 implies that $\mathbf{E} \mid \underline{p}_{jn} - \pi_j \mid^s = \mathcal{O}(n^{-\frac{s}{2}})$ for all $1 \leq s \leq 6$. Our assumptions do *not* imply asymptotic normality, and they do *not* assert that the $\underline{p}_n$ are averages of a sequence of independent indicators. In the multinomial case, of course, the assumptions are true, and we have $V(\pi) = \Pi - \pi\pi'$, where Π is a diagonal matrix with the elements of π along the diagonal.

Definition 2.5. An *estimator* Φ is a function from $\mathbb{S}^m$ to $\mathbb{S}^m$.

Definition 2.5 is somewhat non-standard, because often estimators map data into lower-dimensional (parameter) spaces. Thus $\Phi : \mathbb{S}^m \Rightarrow \mathbb{R}^r$, with $r \leq m$. But in fact this amounts to much the same thing, it merely means that the range of our estimator Φ is a manifold of dimension r. We assume

Assumption 2.6. Φ is three times totally differentiable in a neighborhood of π, and the derivatives are all bounded in that neighborhood.

To measure quality of estimators, we basically look at the distance to the target. Thus we need another definition.

Definition 2.7. A *loss function* Δ is a real-valued function on $\mathbb{S}^m \times \mathbb{S}^m$ satisfying

$$\Delta(x,y) = \begin{cases} \geq 0 & \text{for all } x,y \in \mathbb{S}^m, \\ = 0 & \text{for all } x,y \in \mathbb{S}^m \text{ with } x = y. \end{cases}$$

The function Δ is distance-like, but it need not be symmetric, nor need it satisfy the triangular inequality. For technical purposes we assume

Assumption 2.8. Δ is three times totally differentiable in a neighborhood of $(\pi, \Phi(\pi))$, and the derivatives are bounded in that neighborhood.

In De Leeuw [7] it is assumed that the estimate Φ is a *minimum distance estimate*, i.e. it is computed by minimizing $\Delta(\underline{p}_n, p)$ over p in some sort of *model* Ω, where a model is just a subset of $\mathbb{S}^m$. In the work of Brown and Cudeck, it is assumed, in addition, that the discrepancy function is *appropriate* for the multinomial experiment, in the sense that it gives efficient estimates if the model is true. We do not make any of these assumptions in this paper, in fact we do not even assume that we actually deal with a multinomial or even an asymptotically normal experiment. In our setup, there can be dependence or overdispersion, and some of the higher order moments need not even exist. The only requirement is that our three assumptions 2.2, 2.3, 2.4 are true.

3. Quality of Estimators

We measure the quality of the estimate by using the loss function 2.7 to measure the distance between the target and the estimate.

Definition 3.1. The *overall error* is

$$\underline{\beta}_n \stackrel{\Delta}{=} \Delta(\pi, \Phi(\underline{p}_n)).$$

We also define $\beta_n \stackrel{\Delta}{=} \mathbf{E}(\underline{\beta}_n)$, the *expected overall error* or EOE.

Remark 3.2. Obviously the errors depend on π, but we surpress that dependence in our notation. Linhart and Zucchini calls the EOE the *overall discrepancy*, De Leeuw simply calls it the *bias*.

Now suppose $\underline{q}_n$ is a second sequence of random proportions with the same asymptotic distribution as $\underline{p}_n$, but independent of $\underline{p}_n$. We call $\underline{q}_n$ the *replication*, i.e. it is the outcome of an independent replication of our experiment. Actually, we merely need to assume that Assumptions 2.2, 2.3, and 2.4 are true for $\underline{q}_n$ as well, i.e. the $\underline{q}_n$ are an independent sequence of random variables, converging to the same target as $\underline{p}_n$, with the same speed.

Definition 3.3. The *prediction error* is defined as

$$\underline{\mu}_n \stackrel{\Delta}{=} \Delta(\underline{q}_n, \Phi(\underline{p}_n)).$$

The *expected prediction error* or EPE is, obviously, $\mu_n \stackrel{\Delta}{=} \mathbf{E}(\underline{\mu}_n)$, where expectation is both over $\underline{q}_n$ and $\underline{p}_n$.

Remark 3.4. The EPE is called the *distortion* by De Leeuw. Linhart and Zucchini do not use prediction errors in their work. Cudeck and Browne [5] call the prediction error the *cross validation index*, and give ways to approximate it in [3].

Definition 3.5. The *specification error* is defined as

$$\delta \triangleq \Delta(\pi, \Phi(\pi)).$$

Remark 3.6. Linhart and Zucchini [12] calls this the *discrepancy due to approximation*. The specification error is a non-random quantity, so there is no need for expectations. It also does not vary with n.

Definition 3.7. The *deviance* is

$$\underline{\lambda}_n \triangleq \Delta(\underline{p}_n, \Phi(\underline{p}_n)).$$

And of course we also have the *expected deviance* or EDE $\lambda_n \triangleq \mathbf{E}(\underline{\lambda}_n)$.

Remark 3.8. The deviance does not seem to be a very interesting quantity in itself, but in our interpretation of the estimation situation we only observe $\underline{p}_n$, or at least a realization of $\underline{p}_n$. We do not observe $\underline{q}_n$ or π, and thus the prediction error and specification error cannot be observed directly. There is, of course, a familiar trick to emulate an observable $\underline{q}_n$. If we split our data in two halves, we can use the first half as a realization of $\underline{p}_{n/2}$ and the second half as a realization of $\underline{q}_{n/2}$. We shall use other, Jackknife-based, techniques below, because they seem to be somewhat more systematic.

Definition 3.9. The *estimation error* is defined as

$$\underline{\epsilon}_n \triangleq \Delta(\Phi(\pi), \Phi(\underline{p}_n)).$$

It also has an *expected estimation error* or EEE, which is $\epsilon_n \triangleq \mathbf{E}(\underline{\epsilon}_n)$.

Remark 3.10. Linhart and Zucchini call this the *discrepancy due to estimation*.

Of course generally the various error measures we have defined above do not only have expectations, but also variances, and if suitably normed perhaps even asymptotic distributions. In this paper, as in [7], we concentrate on the expectations.

4. The Distance Geometry of Estimation

A study of the definitions in Section 3 shows that we have six basic quantities between which we measure discrepancies.

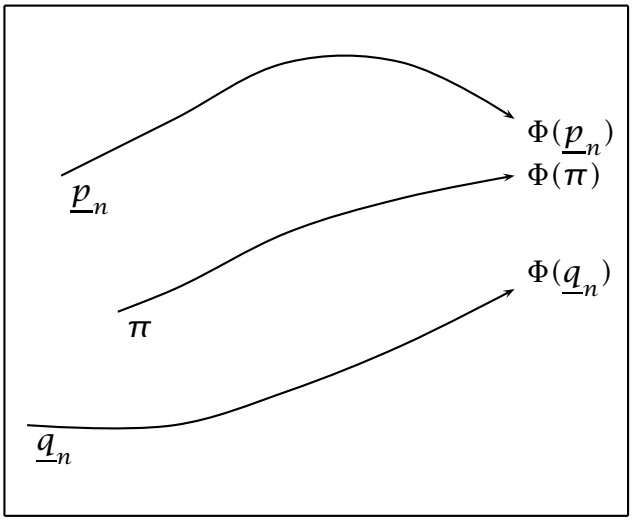

FIGURE 4.1. Geometry of Estimation

There is the target π and its image $\Phi(\pi)$, the data $\underline{p}_n$ and its image $\Phi(\underline{p}_n)$, and the replication $\underline{q}_n$ and its image $\Phi(\underline{q}_n)$. These six basic quantities have 30 nontrivial discrepancies between them. If the loss function is symmetric in its two arguments, then there are still 15 discrepancies. The situation is graphically shown in Figure 4.1.

Some of these 30 discrepancies, and their expected values, were actually named in Section 3, but many remained anonymous. Nevertheless, even the anonymous ones sometimes are of interest. For instance, $\Delta(\underline{p}_n, \underline{q}_n)$ could be used as a measure of dispersion, and it could be compared with $\Delta(\Phi(\underline{p}_n), \Phi(\underline{q}_n))$ to show that estimation enhances precision. Also, as we shall see further on, we can also put Jackknifed or Bootstrapped versions of $\underline{p}_n$ in this same space.

The *idea* behind out setup is that is that $\hat{\pi}_n \stackrel{\Delta}{=} \Phi(\underline{p}_n)$ estimates π, but this notion is not really defined anywhere. In fact, we could say that the *estimation sequence* $\Phi(\underline{p}_n)$ really estimates $\Phi(\pi)$, because obviously

$$\Phi(\underline{p}_n) \stackrel{\mathcal{P}}{\Longrightarrow} \Phi(\pi). \tag{4.1}$$

This is why $\epsilon_n(\pi)$ is called the estimation error. But to a certain extent, this is cheating. Although trivially we are estimating $\Phi(\pi)$, we want to know how good our estimate of $\Phi(\pi)$ is as an estimator of π. To assert that we are estimating what we are estimating, and then by definition consistently, has been referred to as "debilitating relativism".

We can make things a little bit more intuitive by defining the *model* Ω associated with an estimator Φ as the set of fixed points of Φ. Thus

$$\Omega = \{p \in \mathbb{S}^m \mid \Phi(p) = p\}. \tag{4.2}$$

Since $\Phi(\underline{p}_n) \stackrel{\mathcal{P}}{\Longrightarrow} \Phi(\pi)$, we know that if $\pi \in \Omega$, then $\Phi(\underline{p}_n) \stackrel{\mathcal{P}}{\Longrightarrow} \pi$, i.e. we have consistency if π is on the model. This is sometimes expressed as saying that *the model is true*.

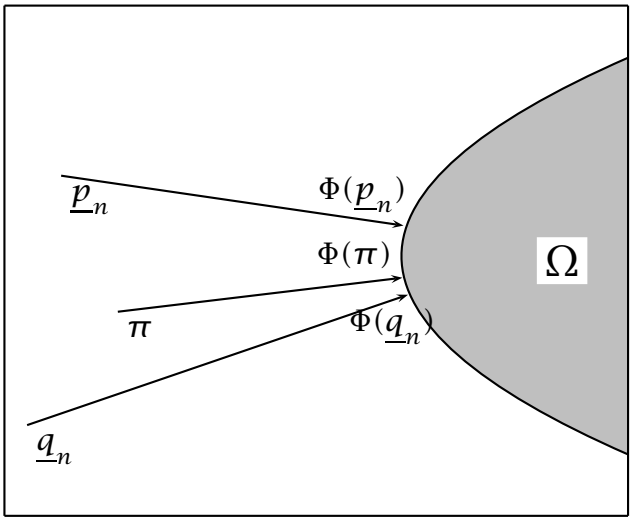

FIGURE 4.2. Model Fitting

On the other hand we can also start with the model Ω, i.e. add it as an ingredient to our estimation problem, and then define Φ to be *F-consistent* for Ω if $\Phi(p) = p$ for all $p \in \Omega$. There is a standard way of guaranteeing F-consistency, which we already indicated in Section 2. We define Φ by the minimum distance or minimum discrepancy principle, which means that we choose $\Phi(\underline{p}_n)$ as the argmin of $\Delta(\underline{p}_n, p)$ over $p \in \Omega$. Or, to put it more geometrically, $\Phi(\underline{p}_n)$ is the *projection* of $\underline{p}_n$ on the model, in the metric defined by the loss function. This is illustrated in Figure 4.2.

For most of this paper, the notion of a model is not needed, however. It suffices to consider the situation where an estimate of π has magically appeared out of the statisticians hat.

5. SOME SIMPLE EXPANSIONS

It is clear from the previous sections that we are interested in statistics of the form $\Delta(F_1(\underline{x}_n), F_2(\underline{y}_n))$, where F_1 and F_2 can be either Φ or the identity, and $\underline{x}_n$ and $\underline{y}_n$ can be either $\underline{p}_n, \underline{q}_n$ or π. Choosing all combinations give a total of $2 \times 2 \times 3 \times 3 = 36$ possibilities, of which 6 give a zero discrepancy because both $F_1 = F_2$ and $\underline{x}_n = \underline{y}_n$.

To derive our basic result, we need some simple definitions. The first partials of the loss function are a

$$t(x, y) \stackrel{\Delta}{=} \mathcal{D}_1\Delta(x, y), \tag{5.1a}$$

$$u(x, y) \stackrel{\Delta}{=} \mathcal{D}_2\Delta(x, y), \tag{5.1b}$$

and second partials are a

$$A(x, y) \stackrel{\Delta}{=} \mathcal{D}_{11}\Delta(x, y), \tag{5.2a}$$

$$B(x, y) \stackrel{\Delta}{=} \mathcal{D}_{12}\Delta(x, y), \tag{5.2b}$$

$$C(x, y) \stackrel{\Delta}{=} \mathcal{D}_{22}\Delta(x, y). \tag{5.2c}$$

We also need a

$$G_1(p) \stackrel{\Delta}{=} \mathcal{D}F_1(p), \tag{5.3a}$$

$$G_2(p) \stackrel{\Delta}{=} \mathcal{D}F_2(p), \tag{5.3b}$$

and a

$$H_1(p) \stackrel{\Delta}{=} \mathcal{D}G_1(p), \tag{5.4a}$$

$$H_2(p) \stackrel{\Delta}{=} \mathcal{D}G_2(p). \tag{5.4b}$$

Observe that for $k = 1, 2$ the function $H_k(\bullet)$ maps $\mathbb{R}^m \otimes \mathbb{R}^m$ into $\mathbb{R}^m$, which means that $H_k(\bullet)(z, z)$ can be written as

$$H_k(\bullet)(z,z) = \begin{bmatrix} z'H_{k1}(\bullet)z \\ \vdots \\ z'H_{km}(\bullet)z \end{bmatrix} \tag{5.5}$$

with each of the $H_{kj}(\bullet)$ square symmetric matrices. The fact that the H_k are not matrices is actually only a minor nuisance, because the quantities we really need are

$$\Gamma_1(x,y) \stackrel{\Delta}{=} \sum_{j=1}^{m} t_j(x,y)H_{1j}(\pi), \tag{5.6a}$$

$$\Gamma_2(x,y) \stackrel{\Delta}{=} \sum_{j=1}^{m} u_j(x,y)H_{2j}(\pi), \tag{5.6b}$$

and these are matrices. We continue the orgy of definitions with

$$V_{11}(\pi) \stackrel{\Delta}{=} n\mathbf{C}\,(\underline{x}_n, \underline{x}_n), \tag{5.7a}$$

$$V_{12}(\pi) \stackrel{\Delta}{=} n\mathbf{C}\,(\underline{x}_n, \underline{y}_n), \tag{5.7b}$$

$$V_{22}(\pi) \stackrel{\Delta}{=} n\mathbf{C}\,(\underline{y}_n, \underline{y}_n), \tag{5.7c}$$

and

$$W_{11}(x,y) \stackrel{\Delta}{=} \Gamma_1(x,y) + G_1(\pi)A(x,y)G_1'(\pi), \tag{5.8a}$$

$$W_{12}(x,y) \stackrel{\Delta}{=} G_1(\pi)B(x,y)G_2'(\pi), \tag{5.8b}$$

$$W_{22}(x,y) \stackrel{\Delta}{=} \Gamma_2(x,y) + G_2(\pi)C(x,y)G_2'(\pi). \tag{5.8c}$$

Finally, $V_{21}(\pi)$ and $W_{21}(x,y)$ are defined by symmetry, and the four submatrices from Equations 5.7a-5.7c and 5.8a-5.8c are collected in matrices $\overline{V}(\pi)$ and $\overline{W}(x,y)$. We are now finally read to state the main result of this section.

Theorem 5.1.

$$2n\mathbf{E}\,(\Delta(F_1(\underline{x}_n), F_2(\underline{y}_n)) - \Delta(F_1(\pi), F_2(\pi))) \Rightarrow \; tr\,\overline{W}(F_1(\pi), F_2(\pi))\overline{V}(\pi).$$

Proof. Apply the first half of Hurt's Theorem, from the Appendix, with $q = 2$. □

6. Estimating the Various Errors

We now apply Theorem 5.1 to some of the expected error measures we have defined earlier. Unfortunately, we need more definitions. Let

$$G(r) \triangleq \left.\frac{\partial \Phi}{\partial p}\right|_{p=r}, \tag{6.1}$$

Corollary 6.1. *For the expected overall error*

$$2n(\beta_n - \delta) \Rightarrow \ tr\,[\Gamma(\pi, \Phi(\pi)) + G(\pi)C(\pi, \Phi(\pi))G'(\pi)]V(\pi).$$

Proof. Here F_1 is the identity and $F_2 = \Phi$. Also $\underline{x}_n = \pi$, and $\underline{y}_n = \underline{p}_n$. □

Corollary 6.2. *For the expected prediction error*

$$2n(\mu_n - \delta) \Rightarrow \ tr\,[A(\pi, \Phi(\pi)) + \Gamma(\pi, \Phi(\pi)) + G(\pi)C(\pi, \Phi(\pi))G'(\pi)]V(\pi)$$

Proof. Here F_1 is the identity, $F_2 = \Phi$, $\underline{x}_n = \underline{q}_n$, and $\underline{y}_n = \underline{p}_n$. □

Corollary 6.3. *For the expected deviance*

$$2n(\lambda_n - \delta) \Rightarrow \ tr\,[A(\pi, \Phi(\pi)) + B(\pi, \Phi(\pi))G'(\pi) + G(\pi)B'(\pi, \Phi(\pi)) + \Gamma(\pi, \Phi(\pi)) + G(\pi)C(\pi, \Phi(\pi))G'(\pi)]V(\pi)$$

Proof. Here F_1 is the identity, $F_2 = \Phi$, and $\underline{x}_n = \underline{y}_n = \underline{p}_n$. □

Corollary 6.4. *For the expected estimation error*

$$2n\epsilon_n \Rightarrow \ tr\,[G(\pi)C(\Phi(\pi), \Phi(\pi))G'(\pi)]V(\pi).$$

Proof. Here $F_1 = F_2 = \Phi$, $\underline{x}_n = \pi$, and $\underline{y}_n = \underline{p}_n$. Moreover we use $\Gamma_2(\Phi(\pi), \Phi(\pi)) = 0$, because $u_j(\Phi(\pi), \Phi(\pi)) = 0$ for all $j = 1, \cdots, m$. □

The formula's in the corrollaries are a bit hard to use, partly because they involve the complicated matrix Γ of second derivatives of the estimator. As an example, we give the following result.

Theorem 6.5. *Define*

$$2n\underline{\mathcal{D}}_n \triangleq 2n\underline{\lambda}_n - \ tr\,[A(\underline{p}, \Phi(\underline{p})) + B(\underline{p}, \Phi(\underline{p}))G'(\underline{p}) + G(\underline{p})B'(\underline{p}, \Phi(\underline{p})) + \Gamma(\underline{p}, \Phi(\underline{p})) + G(\underline{p})C(\underline{p}, \Phi(\underline{p}))G'(\underline{p})]V(\underline{p}).$$

Then

$$n\mathbf{E}\,(\underline{\mathcal{D}}_n) \Rightarrow \delta.$$

Proof. Direct from Corollary 6.3. □

Another problem is that the formulas require values for some of the unobserved quantities. But, as explained in De Leeuw [7], there is a simple and fairly straightforward way around this problem.

Theorem 6.6. *Define*

$$2n\underline{\mathcal{P}}_n \triangleq 2n\underline{\lambda}_n - \ tr\,[B(\underline{p}_n, \Phi(\underline{p}_n))G'(\underline{p}_n) + G(\underline{p}_n)B'(\underline{p}_n, \Phi(\underline{p}_n))]V(\underline{p}_n),$$

$$2n\underline{\mathcal{B}}_n \triangleq 2n\underline{\lambda}_n - \ tr\,[A(\underline{p}_n, \Phi(\underline{p}_n)) + B(\underline{p}_n, \Phi(\underline{p}_n))G'(\underline{p}_n) + G(\underline{p}_n)B'(\underline{p}_n, \Phi(\underline{p}_n))]V(\underline{p}_n).$$

Then

$$\mathbf{E}(\underline{\mathcal{P}}_n) \Rightarrow \mu_n,$$
$$\mathbf{E}(\underline{\mathcal{B}}_n) \Rightarrow \beta_n.$$

Proof. From Corollaries 6.2 and 6.1. □

7. Moment Structure Analysis

We now generalize our results to statistics of the form

$$\underline{x}_n = \Psi(L\underline{p}_n). \tag{7.1}$$

Here $\underline{p}_n$ are $m-$vectors of proportions, as before, L is a known $p \times m$ matrix, and Ψ maps $\mathbb{R}^p$ into $\mathbb{R}^s$. We estimate $\underline{x}_n$ by functions of the form $\Phi(\underline{x}_n)$, which map $\mathbb{R}^s$ into $\mathbb{R}^s$. Correspondingly, we define a

$$\mu \stackrel{\Delta}{=} \Psi(L\pi), \tag{7.2a}$$

$$\underline{y}_n \stackrel{\Delta}{=} \Psi(L\underline{q}_n). \tag{7.2b}$$

With these definitions, our previous results and definitions remain valid, with $\underline{x}_n$ substituted for $\underline{p}_n$, $\underline{y}_n$ substituted for $\underline{q}_n$, and μ substituted for π.

To indicate the generality of this situation, think of a p-dimensional random variable $\underline{\ell}$ which takes the values $\ell_1, \cdots \ell_m$ with probabilities $\pi_1, \cdots, \pi_m$. The ℓ_j are the columns of L. Then $\mu = L\pi$ is the expectation of $\underline{\ell}$, and n replications of $\underline{\ell}$ bring us to the $\underline{x}_n$ of 7.1. But this is not all.

If $\underline{z}$ takes the scalar values $z_1, \cdots, z_m$, with probabilities $\pi_1, \cdots, \pi_m$, then $\underline{\ell} = (\underline{z}, \underline{z}^2, \cdots, \underline{z}^p)$ takes on values $\ell_j = (z_j, z_j^2, \cdots, z_j^p)$ with the same probabilities, and the $\underline{x}_n$ are the first p sample moments. Thus the theory covers functions of the sample moments.

A slightly more general setup extends it to functions of sample moments and product moments, and functions of suitably discreticized versions of the empirical distribution function or the empirical characteristic function. The usual analysis of covariance structures is covered painlessly, because it concerns itself with functions of the second order product moments (and, in the case of asymptotically distribution-free methods, functions of the product moments up to order four, cf [6]).

8. Use of the Jackknife

Although Theorems 6.5 and 6.6 can be applied in considerable generality, they do require quite heavy derivative-type calculations. This may not seem to be a large disadvantage, since differentiation is usually straightforward, although tedious. The major problem, however, is that automating differentiation is not simple. Unless sophisticated symbolic mathematics packages are used, differentiation has to be done by hand, and then programmed into subroutines which are problem-specific.

Alternatively, we can do numerical differentiation. This is automatic, and comparatively easy to program in full generality, but we still have to make some strategic choices on how to implement it precisely. In this paper we use the Jackknife for our numerical differentiation, because it is specifically designed for statistical problems, and in particular for statistical problems dealing with proportions.

Jackknife-based techniques are somewhat less general than the similar delta-method expansions, because they actually suppose a strictly multinomial model. I am sure they can be adapted quite easily, but here we simply assume that $V(\pi)=\Pi-\pi\pi'$. If we want to use the Jackknife, we use perturbations in our differentiation of the form

$$p_{n:j}\overset{\Delta}{=}p_n+\frac{1}{n-1}(p_n-e_j), \tag{8.1}$$

where the e_j are the unit-vectors in $\mathbb{R}^m$. Thus

$$p_n=\sum_{j=1}^m p_{nj}e_j. \tag{8.2}$$

We also set

$$V(p_n)\overset{\Delta}{=}\sum_{j=1}^m p_{nj}(p_n-e_j)(p_n-e_j)'=P_n-p_np_n', \tag{8.3}$$

with P_n the diagonal matrix with the p_n. Define the following functions. They are defined to mimic the expansions of the error measures we have derived in the previous sections.

Definition 8.1.

$$\mathcal{J}_n(p_n)\overset{\Delta}{=}2(n-1)^2\{\sum_{j=1}^m p_{nj}\Delta(p_{n:j},\Phi(p_n))-\Delta(p_n,\Phi(p_n))\},$$

$$\mathcal{K}_n(p_n)\overset{\Delta}{=}2(n-1)^2\{\sum_{j=1}^m p_{nj}\Delta(p_n,\Phi(p_{n:j}))-\Delta(p_n,\Phi(p_n))\},$$

$$\mathcal{L}_n(p_n)\overset{\Delta}{=}2(n-1)^2\{\sum_{j=1}^m p_{nj}\Delta(p_{n:j},\Phi(p_{n:j}))-\Delta(p_n,\Phi(p_n))\}.$$

Lemma 8.2.

$$\begin{aligned}\mathcal{J}_n(p_n)&\Rightarrow\ tr\,A(p_n,\Phi(p_n))V(p_n),\\ \mathcal{K}_n(p_n)&\Rightarrow\ tr\,[\Gamma(p_n,\Phi(p_n))+G(p_n)C(p_n,\Phi(p_n))G'(p_n)]V(p_n),\end{aligned}$$

and

$$\begin{aligned}\mathcal{L}_n(p_n)\Rightarrow\ tr\,[&A(p_n,\Phi(p_n))+\Gamma(p_n,\Phi(p_n))+\\ &G(p_n)B(p_n,\Phi(p_n))+B(p_n,\Phi(p_n))G'(p_n)+\\ &G(p_n)C(p_n,\Phi(p_n))G'(p_n)]V(p_n).\end{aligned}$$

Proof. These are just simple expansions. □

Theorem 8.3.

$$\mathbf{E}\,(2n\underline{\lambda}_n-\mathcal{L}_n(\underline{p}_n))\Rightarrow\delta$$

Proof. From Lemma 8.2 and Theorem 6.5. □

Theorem 8.4.

$$\begin{aligned}\mathbf{E}\,(2n\underline{\lambda}_n-\{\mathcal{L}_n(\underline{p}_n)-\mathcal{K}_n(\underline{p}_n)\})&\Rightarrow\beta_n,\\ \mathbf{E}\,(2n\underline{\lambda}_n-\{\mathcal{J}_n(\underline{p}_n)+\mathcal{K}_n(\underline{p}_n)-\mathcal{L}_n(\underline{p}_n)\})&\Rightarrow\mu_n.\end{aligned}$$

Proof. From Lemma 8.2 and Theorem 6.6. □

It is clear from the form of Theorems 8.3 and 8.4, and the type of reasoning that is used, that they generalize without modification to the more general situation explained in Section 7.

9. Discussion

It is quite obvious what the computational costs are of the methods in Theorems 8.3 and 8.4. In stead of computing $\Phi(p_n)$ once, we have to compute it the m values $\Phi(p_{n:j})$. If m is larger than n, we can also develop a version which recomputes the estimate n times (leave out one observation at a time). For complicated estimates, such as those generated by LISREL or EQS runs, and for large samples this may be a problem (although we can start the iterative process to compute the estimates in $\Phi(p_n)$, which will in general be quite close to the perturbed estimate). It is also clear that for large samples, we need pretty high precision in our computations, otherwise the effect of the perturbations will get lost in the rounding errors.

If the Jackknife based computations are inpractical, there are some alternatives. We can use a random version of the Jackknife, which leaves out cells or observations at random. Or we can use the expansions in Theorems 6.5 and 6.6. Or we can use the Bootstrap (as tried in De Leeuw [7]).

I prefer these computational techniques to techniques that make very specific assumptions. Some of these assumptions have been mentioned above. We can assume that the model is true, i.e. that $\Phi(\pi) = \pi$, or that the estimates are minimum distance, or that the loss functions are "appropriate", or even that we use likelihood methods throughout. I see no compelling reason for making any of these assumptions in many cases, and I can think of many reasons for not making them.

REFERENCES

1. P. M. Bentler, *Comparative Fit Indices in Structural Models*, Psychological Bulletin **107** (1990), 238–246.
2. P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models*, John Hopkins University Press, Baltimore, MD, 1993.
3. M. W. Browne and R. Cudeck, *Single Sample Cross-Validation Inidices for Covariance Structures*, Multivariate Behavioral Research **24** (1989), 445–455.
4. _____, *Alternative Ways of Assessing Model Fit*, Sociological Methods and Research **21** (1992), 230–258.
5. R. Cudeck and M. W. Browne, *Cross-Validation of Covariance Structures*, Multivariate Behavioral Research **18** (1983), 147–167.
6. J. de Leeuw, *Models and Methods for the Analysis of Correlation Coefficients*, Journal of Econometrics **22** (1983), 113–137.
7. _____, *Model Selection in Multinomial Experiments*, On Model Uncertainty and its Statistical Implications (T. J. Dijkstra, ed.), Springer Verlag, New York, NY, 1988.
8. A.E. Gelfand, D.K. Dey, and H. Chang, *Model Distributions using Predictive Distributions with Implementation via Sampling-Based Methods*, Bayesian Statistics 4, Oxford University Press, New York, NY, 1992.
9. J. Hurt, *Asymptotic Expansions of Functions of Statistics*, Aplikace Matematiky **21** (1976), 444–456.
10. _____, *Asymptotic Expansions for Moments of Functions of Statistics*, Second Prague Symposium on Asymptotic Statistics, Akademie, Prague, Czechoslovakia, 1978.
11. B. Lindsay, *Efficiency versus Robustness: The Case for Minimum Hellinger Distance and Related methods*, Annals of Statistics **22** (1994), 1081–1114.
12. H. Linhart and W. Zucchini, *Model Selection*, Wiley, New York, NY, 1986.
13. R. P. McDonald, *An Index of Goodness-of-Fit Based on Noncentrality*, Journal of Classification **6** (1989), 97–103.
14. R. P. McDonald and H. W. Marsh, *Choosing a Multivariate Model: Noncentrality and Goodness of Fit*, Psychological Bulletin **107** (1990), 247–255.

APPENDIX A. HURT'S THEOREM

In this Appendix we give a relevant theorem that can be used to justify the approximations to the expected values used in the paper. It is, to some extent, classical, but a particulary clear and useful statement appears in two publications by Jan Hurt [9, 10].

Theorem A.1. *Suppose*

(1) $\{\phi_n(x)\}$ *is a sequence of real-valued functions on* $\mathbb{R}^m$,
(2) $\{\underline{x}_n\}=\{(\underline{x}_{1n},\cdots\underline{x}_{mn})\}$ *is a sequence of* $m-$*dimensional statistics.*

Assume that

(1) *For all* n, ϕ_n *is* $(q+1)$*-times totally differentiable with respect to the* x_j *in the interval* $\mathbb{K}=\prod_{j=1}^m[\mu_j-\delta_j,\mu_j+\delta_j],\delta_j>0,\delta_j$ *independent of* n.
(2) *For all* n, ϕ_n *is bounded on* $\mathbb{R}^m$.
(3) *For all* n, *the derivatives* $\phi_n^{(1)},\cdots,\phi_n^{(q+1)}$ *are bounded on* $\mathbb{K}$.
(4) *For all* j *and all* n, $\underline{x}_{jn}$ *has finite absolute moments up to the order* $2(q+1)$.
(5) *For all* $j=1,\cdots,m$

$$\boldsymbol{E}\mid\underline{x}_{jn}-\mu_j\mid^{2(q+1)}=\mathcal{O}(n^{-(q+1)}).$$

Then

$$\mathbf{E}\left(\phi_n(\underline{x}_n)-\phi_n(\mu)\right)=\sum_{j=1}^{q}\frac{1}{q!}\sum_{\substack{\mathrm{i}_1+\cdots+\mathrm{i}_\mathrm{m}=\mathrm{j}\\ i_1,\cdots,i_m\geq 0}}\cdots\sum\left(\frac{\partial^j\phi_n}{\partial x_1^{i_1}\cdots\partial x_m^{i_m}}\right)_{x=\mu}\times$$
$$\times\,\mathbf{E}\left((\underline{x}_{1n}-\mu_1)^{i_1}\cdots(\underline{x}_{mn}-\mu_m)^{i_m}\right)+\mathcal{O}(n^{-\frac{(q+1)}{2}}).$$

and

$$\mathbf{V}\left(\phi_n(\underline{x}_n)-\phi_n(\mu)\right)=\sum_{\substack{\mathrm{j}=1\\ j+k\leq q+1}}^{\mathrm{q}}\sum_{\mathrm{k}=1}^{\mathrm{q}}\frac{1}{j!}\frac{1}{k!}\sum_{\substack{\mathrm{i}_1+\cdots+\mathrm{i}_\mathrm{m}=\mathrm{j}\\ i_1,\cdots,i_m\geq 0}}\cdots\sum\sum_{\substack{\ell_1+\cdots+\ell_\mathrm{m}=\mathrm{k}\\ \ell_1,\cdots,\ell_m\geq 0}}\cdots\sum\times$$
$$\times\left(\frac{\partial^j\phi_n}{\partial x_1^{i_1}\cdots\partial x_m^{i_m}}\right)_{x=\mu}\left(\frac{\partial^k\phi_n}{\partial x_1^{\ell_1}\cdots\partial x_m^{\ell_m}}\right)_{x=\mu}\times$$
$$\times\,\mathbf{C}\left((\underline{x}_{1n}-\mu_1)^{i_1}\cdots(\underline{x}_{mn}-\mu_m)^{i_m},(\underline{x}_{1n}-\mu_1)^{i_1}\cdots(\underline{x}_{mn}-\mu_m)^{i_m}\right)+$$
$$+\,\mathcal{O}(n^{-\frac{(q+2)}{2}}).$$

Proof. See Hurt [9]. □

Remark A.2. In [10] the same result is proved, but the residuals are not expressed in terms of powers of n, but in terms of the absolute moments. This extends the result to a more general class of statistics.

UCLA STATISTICS PROGRAM, 8118 MATHEMATICAL SCIENCES BUILDING, UNIVERSITY OF CALIFORNIA AT LOS ANGELES
E-mail address: deleeuw@stat.ucla.edu