A Monte Carlo Comparison of Methods for Compositional Data Analysis

John Brehm, Duke University
Scott Gates, Michigan State University
Brad Gomez, Duke University

July 8, 1998

Abstract

This paper offers an explication of two techniques for compositional data analysis, which involve non-negative data belonging to mutually exclusive and exhaustive categories. The Dirichlet distribution is a multivariate generalization of the beta distribution that offers considerable flexibility, and ease of use, but requires a strong form of an "independence of irrelevant alternatives" (IIA) assumption. The second application, proposed by Aitchison (1986) and applied to political data by Katz and King (1997), is the additive logistic method. This approach addresses the strong IIA assumption, but cannot handle strong forms of independence (Rayens and Srinivasen 1994). Monte Carlo simulations are employed on compositional data to explore the limits of applications of the two methods. Data on police officers' allocation of time across a variety of tasks (Ostrom et al. 1988) is used in this analysis. Comparing both common covariates and unique covariates. When the composites are influenced by common covariates, there appears to be no advantage in the use of additive logistic methods over the Dirichlet. Similarly, the additive logistic and Dirichlet methods appear to be equally successful at estimating the effects of the unique covariates on composites. From these simulation results we conclude that the additive logistic method offers little advantage over the Dirichlet, and suffers from several disadvantages.

Prepared for presentation at the 1998 annual meeting of the Society for Political Methodology. Thanks to Jeff Gill, Jonathan Katz, and Gary King for consultation.

1 Compositional Data Analysis

Compositional data analysis offers the opportunity for new approaches to problems of long-standing importance in the social sciences. The method refers to analysis of non-negative data referring to mutually exclusive and exhaustive categories, where the data sum to one. The constraints thus imposed on the data mean that analysis of any one category must take into account the balance in the other categories: an increase in one category must be compensated by a decrease distributed across all the other categories. That is, for each observation i within J composites.

$$y_{ij} > 0, \ \forall j = 1 \ldots J \tag{1}$$

$$\sum_{j=1}^{J} y_{ij} = 1. \tag{2}$$

These two features mean that the composites y_{ij} constitute a *simplex.* Mathematical features of a simplex will mean that the distributions of the y_{ij} are not fully independent. For example, if one knows the values of y_{i1} up through $y_{i(J-1)}$, then one knows the value of y_{iJ}.

This paper offers an explication of two techniques for compositional data analysis. One technique (in more common use prior to 1986) is application of maximum likelihood estimates of Dirichlet densities; the other technique, popularized by John Aitchison in 1986, is application of seemingly unrelated regression-like analysis of log-ratios of composites. Ultimately, the purpose of this paper is to employ Monte Carlo simulations of compositional data taking different archetypal forms to explore the limits of applications of the two methods. The paper begins with further substantive motivation behind use of the methods; second, turns to an elaboration upon the two approaches; and third, provides details about the Monte Carlo simulations and the implications of simulation results.

Jonathan Katz and Gary King have recently developed and applied one class of methods for compositional data analysis (maximum likelihood estimates of the additive logistic multivariate-t density) to multiparty election results in the U.K. (1997). The proportion of each district's vote that divides across the Conservative, Labour, and Social Democratic parties is an excellent example of compositional data of interest to political scientists. (Note that this is not the same as the multi-candidate choice problem at the individual level, since the unit of analysis is the constituency (or other aggregate); nor is it an ecological inference problem, since the inferences are about the effect of constituency wide variables, such as incumbency, on the aggregate).

There are other appropriate applications of compositional data analysis. Municipal, state, or federal

budgets may be another application, as long as the analysis is of the percentage of the total budget devoted to particular budgetary lines. (Padgett (1981) offers an analysis of budgets which takes on the mathematical form of compositional data analysis, although he does not label it as such. Padgett's application is the first application of the Dirichlet density in political science, to the best of our knowledge.) The percentage of vote splits by ethnic or racial categories (as long as they are mutually exclusive and exhaustive) would be another potential application.

The present substantive motivation is to analyze time budgets for public bureaucrats. In our prior substantive explorations (Brehm and Gates 1997) of the behavior of public bureaucrats, we consider how much effort (or time) bureaucrats devote to working relative to shirking. Our previous analysis demonstrated that supervisors in public bureaucracies can influence the total amount of time or effort devoted by their subordinates, through their use of greater monitoring or application of rewards and sanctions. However, such coercive aspects of supervision were weaker influences on bureaucratic effort than intersubordinate influences, pressures from "customers" (those receiving the services), and (most importantly) the recruitment process into the bureaucracy. Using the language of the economics of the firm, we found that adverse selection factors outweigh those associated with moral hazard.

Supervisors might have much greater effect upon how bureaucrats allocate their time across different forms of working than upon the division of time between working and shirking. That is, suppose that there are three tasks, A, B, and C. A and B are equally liked by the subordinate, and preferred to C ($A = B > C$). The supervisor, however, prefers $C > B > A$. The supervisor acting as a coordinator will probably be more successful in obtaining greater effort from the bureaucrat to task B than in encouraging effort to task C.

One of the essential duties of a supervisor is to allocate tasks across his or her subordinates. Supervisors in this way serve as coordinators of an organization's bureaucratic input. We presume that supervisors will match tasks to subordinates in a manner that maximizes the production of an ideal policy mix for the supervisor. In this manner we see a more complicated interaction between subordinates and superiors than is typically modeled in principal-agent frameworks.

1.1 Police Data

Some specific data at hand come from the 1977 Police Services Study, conducted by Elinor Ostrom, Roger Parks, and Gordon Whittaker in three cities (Rochester, St. Louis, and St. Petersburg). The study combined multiple methods, including observations of police officers' behavior during their shifts. The observational

data provides an excellent opportunity to test our propositions about the allocation of time across tasks. At the conclusion of each shift, the observer recorded the amount of time officers spent on a total of eleven tasks (italicized phrase or word denotes our label in subsequent graphs):

1. Time on administrative duties *(Adm);*
2. Time report writing *(Rept);*
3. Time out of car for foot patrol (not on an encounter or dispatched run) *(Foot Pat);*
4. Time on routine mobile patrol *(Mob Pat);*
5. Time at or en route to an encounter or dispatched run *(Run);*
6. Time on mobile traffic work (radar, vascar, etc.) *(Mob Traf);*
7. Time on stationary traffic work (radar, etc.) *(Stat Traf)*;
8. Time on meals, other 10-7 breaks *(Meal);*
9. Time on mobile personal business *(Mob Pers);*
10. Time on stationary personal business *(Stat Pers);*
11. Time on other stationary police work (surveillance, stake out, etc.) *(Other);*

We apply two different means for examining the amount of time an officer devotes to different tasks. The first of these is the "ternary" diagram, and is most useful when one collapses the distribution of time across tasks into three categories. Here, we consider time spent on personal business (time on meals, stationary and mobile personal business), time completing paperwork (administration, reports), and time policing (mobile and stationary traffic, runs, mobile and foot patrol, and other). The collapsing of time into three tasks corresponds nicely with a division into a police officer's principal responsibilities (policing and paperwork), plus a category denoting time not devoted to official responsibilities (personal). In our previous analysis (Brehm and Gates 1993, 1997), we facetiously referred to these as "donut shops" (shirking, here measured as time on personal business) and "speed traps" (working). In the present analysis, we divide time spent working between the categories of policing ("speed traps") and paperwork.

If the amount of time spent on tasks is transformed into percentages of total time, and total time is constrained to sum to 1, then the data are arranged on the simplex. One could produce a three-dimensional scatterplot of the data across the three dimensions of tasks, but all of the points would fall on the triangular plane intersecting the three axes at 1.0 (Figure 1). Instead, we focus solely on that triangular plane, displayed in Figure 2.

The figure makes it quite clear that the majority of these officers' time is devoted to policing. The mode of the distribution is quite close to the extreme lower right corner, although there is a fair amount of dispersion

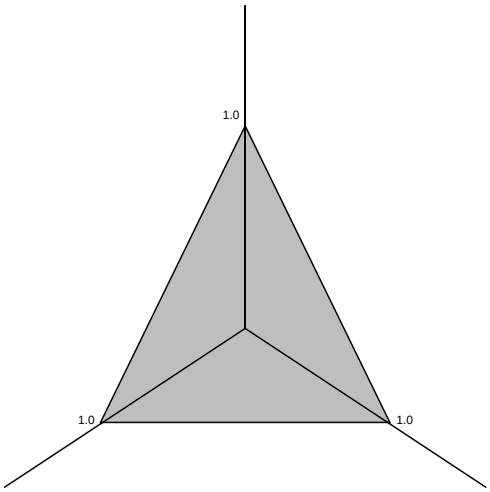

Figure 1: Simplex for Three Dimensions

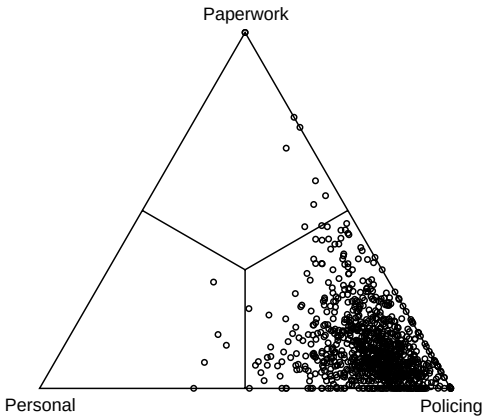

Figure 2: Actual Distribution of Time on Simplex, 1977 Police Data

throughout the lower right trident of the ternary diagram. Only five officers spent a plurality of their time on personal business, running counter to stereotypes about police behavior. Eight officers devoted a plurality of their time to paperwork, including one officer who spent the entire shift on paperwork. There are also some interesting edge conditions — officers who divided their time between either policing and paperwork, or policing and personal business.

The second graphical display (Figure 3) involves use of a technique called the "checkerboard plot." Each officer is displayed as a vertical column of rectangles (here, quite thin—nearly lines—since we need to display over 900 officers' shifts). Each row of rectangles corresponds to one of the 11 tasks (e.g., mobile patrol, meals). We shade each rectangle with a percentage gray to denote amount of time by the officer at that task: rectangles which are completely white denote those tasks where an officer spent zero time at the task, rectangles which are completely black denote those tasks to which the officer devoted his or her entire shift, while those which are gray denote those tasks to which the officer spent some middling fraction. The darker the gray, the more time devoted to the task.

As is readily apparent from the checkerboard plot, police officers spend the majority of their day confined to two tasks: mobile patrol, and on route to an encounter. Officers spend the least amount of their time on foot patrol, mobile and stationary traffic. Officers spend middling amounts of time completing reports or performing other administrative duties, as well as on meals or stationary personal business. (The meals

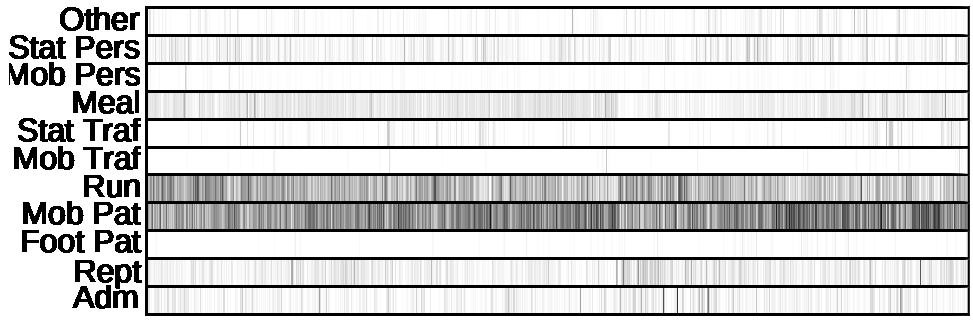

Figure 3: Checkerboard Plot of Actual Distribution of Time on Simplex, 1977 Police Data

category is in third place, on average, although distantly behind runs and mobile patrol).

As is also apparent, these patterns are strikingly homogeneous across the more than 900 police officers in the three different cities. Although one can identify individuals who devote a plurality of time to administration and reports (the dark lines in those sections of the plot), as well as those who engage in nearly twice as much time at meals as other officers, the general pattern here is one of uniformity, not variation.

The substantive problem that motivates the present research is to identify the reasons why the officers would devote more or less time out of their shifts to the different distributions of tasks. In turn, understanding how street-level bureaucrats such as police officers allocate their time across tasks helps us better understand the factors that affect the policy choices of bureaucrats. Although we are rather limited in our set of explanatory variables, there are several that prove illuminating: number of contacts with the supervisor, number of contacts with fellow officers, expressed likes (or dislikes) about colleagues, and expressed likes (or dislikes) about the job. The idea would be to model the systematic component of an appropriate density, and estimate the effect of changes in this (limited) set of variables on the components.

The purpose of this paper, however, is to compare two distinct methods for compositional data analysis. (A separate paper (Brehm, Gates, and Gomez 1998) elaborates on our model of subordinate officers' time budgets, and our estimates of one such approach to compositional data).

2 Two Methods for Compositional Data Analysis

The two methods that we consider here are based upon fundamentally different probability processes (and, hence, different densities). The first method, application of the Dirichlet distribution, is a multivariate generalization of the beta distribution, and offers considerable flexibility, is rapidly estimated with conventional statistical software, but requires a strong form of an "independence of irrelevant alternatives" assumption. The second method, transformation of the composites into log-ratios and estimation via seemingly-unrelated regression-like analysis, offers less flexibility, is estimated less rapidly than the Dirichlet, does not require the strong IIA assumption, but cannot handle strong forms of independence (Rayens and Srinivasan 1994). We discuss the details of each method in turn.

2.1 Dirichlet

One relatively simple solution begins from an assumption that each composite is produced by an independent process. Suppose that y_{ij} is distributed as J independent gamma random variates with shape parameters $\nu_1 \ldots \nu_J$, where the y_{ij} are distributed on the simplex. The composites are then distributed according to a *Dirichlet distribution*:

$$\begin{aligned} (y_1 \ldots y_J) &= f_D(Y_1 \ldots Y_J | \nu_1 \ldots \nu_J) && (3) \\ &= \frac{, (\sum_{k=0}^{J} \nu_k)}{\prod_{k=0}^{J} , (\nu_k)} \prod_{k=1}^{J} y_k^{\nu_k - 1} && (4) \end{aligned}$$

where

$$\nu_j > 0, \forall j = 1 \ldots J \quad (5)$$

One can reparameterize the ν_j in terms of explanatory variables and coefficients with simple exponentiation:

$$\nu_j = \exp(X\beta_j), \quad (6)$$

where the effect parameters (β_j) vary by composite, and the X may or may not be the same set of explanatory variables (identification for the system is accomplished through covariance restrictions, detailed below, and through functional form). If one assumes that the observations are distributed identically and independently, then the log-likelihood for the reparameterized Dirichlet is:

$$\ln L(\beta | X, y) = \sum_{i=1}^{N} \left[\ln , (\sum_{j=1}^{J} e^{X\beta_j}) + \sum_{j=1}^{J} e^{X\beta_j} \ln y_j - \sum_{j=1}^{J} \ln , (e^{X\beta_j}) \right]. \quad (7)$$

This log-likelihood is easily optimized with a package such as Gauss.

Several features of the Dirichlet lend itself to some desirable properties for purposes of interpretation. The Dirichlet is a multivariate generalization of the Beta distribution (which we use extensively in our analysis of the allocation of time across two "tasks," (working and shirking) in Brehm and Gates (1997)). As such, it is a highly flexible distribution permitting multiple modes and asymmetry. Further, the moments are easily found. Let $\nu^* = \sum_{k=1}^{J} \nu_k$. The mean of each composite j is

$$\mu_j = \frac{\nu_j}{\nu^*}. \tag{8}$$

The variance of composite j is

$$\text{var}(y_j) = \frac{\nu_j(\nu^* - \nu_j)}{\nu^{*2}(\nu^* + 1)} \tag{9}$$

and the covariance of composites k and m is

$$\text{cov}(y_k, y_m) = \frac{-\nu_k \nu_m}{\nu^{*2}(\nu^* + 1)}. \tag{10}$$

Since all the ν_j are positive, this means that the covariance of any pair of composites k and m is negative, or that any increase (decrease) in one composite necessitates a decrease (increase) in *every other* composite.

This property of the Dirichlet distribution is the first sign that there are hidden assumptions in the Dirichlet that may warrant another selection of distributional assumptions. Aitchison (1986) writes

> It is thus clear that every Dirichlet composition has a very strong implied independence structure and so the Dirichlet class is unlikely to be of any great use for describing compositions whose components have even weak forms of dependence. ... This independence property, which holds for every partitition of every Dirichlet composition, is again extremely strong, and unlikely to be possessed by many compositions in practice. For example, one implication of it is that each ratio x_i/x_j of two components is independent of any other ratio x_k/x_l formed from two other components.

What remains to be seen, however, is just how sensitive the analysis of composite data is to this particular "strong" IIA (independence of irrelevant alternatives) assumption.

The irony is that the Dirichlet distribution, like the beta distribution, is capable of considerable variation in potential distributions of allocation of the compositions. Figures 4–7 demonstrate simulated Dirichlet distributions for varying selections of the parameters. It is possible to generate, among other forms, Dirichlet which are uniformly dispersed (Figure 4), unimodal and centered (Figure 5), unimodel and off-centered (Figure 6), or multimodal and skewed (Figure 7).

2.2 Additive Logistic

Aitchison and Shen (1980) and Aitchison (1986) offer an alternative to the Dirichlet. One relatively simple method begins from an assumption that each composite is produced by an independent process. Suppose

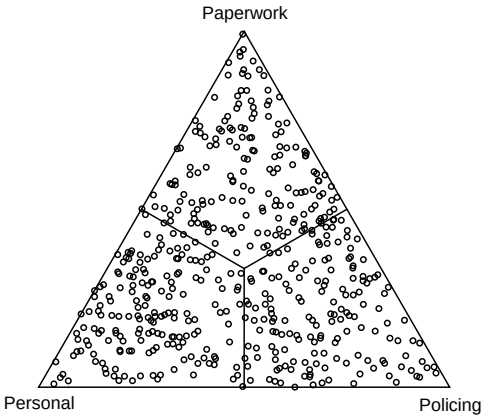

Figure 4: Simulated Dirichlet, $\nu_1 = 1, \nu_2 = 1, \nu_3 = 1$

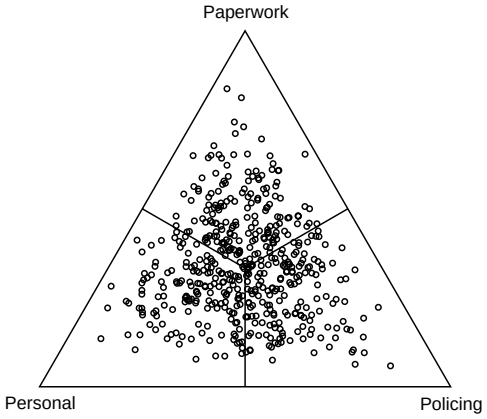

Figure 5: Simulated Dirichlet, $\nu_1 = 3.5, \nu_2 = 3.5, \nu_3 = 3.5$

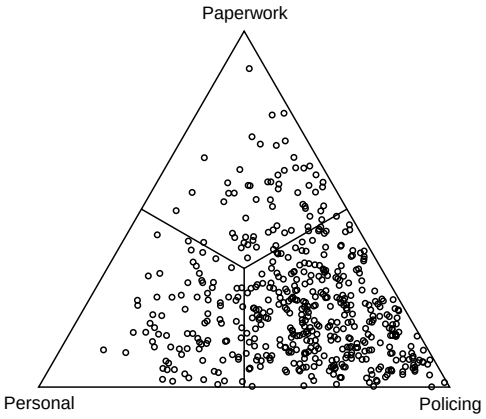

Figure 6: Simulated Dirichlet, $\nu_1 = 1.5, \nu_2 = 1.5, \nu_3 = 3.5$

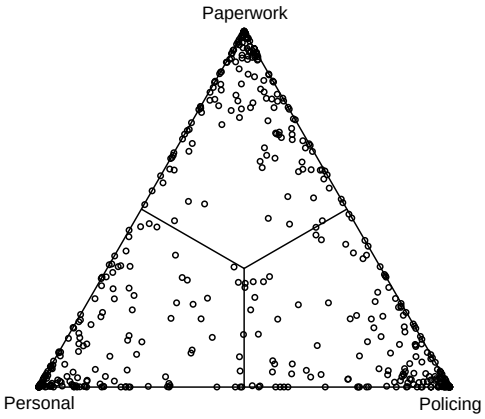

Figure 7: Simulated Dirichlet, $\nu_1 = .25, \nu_2 = .25, \nu_3 = .25$

y_{ij} is distributed as J independent gamma random variates with shape parameters $\nu_1 \ldots \nu_J$. Then the set of composites y_{ij}. Transform the y_{ij} by the familiar log-ratio function relative to a baseline composite

$$v_j = \ln(y_j/y_J), j = 1 \ldots J-1. \tag{11}$$

The v_j are now unbounded. By virtue of this additive logistic transformation relative to the baseline (here, v_J), the v_j are also unconstrained. The relation from the v_j back towards the y_j may be solved through the additive logistic transformation

$$y_j = \exp(v_j)/(\exp(v_1) + \ldots + \exp(v_{J-1}) + 1). \tag{12}$$

If one assumes that the v_j are distributed multivariate normally with mean μ and covariance matrix Σ, then the probability for any distribution of v_i is

$$\begin{aligned} \Pr(v_i|\mu,\Sigma) &= (2\pi)^{-3/2}||\Sigma||^{-1/2}(v_1 v_2 \ldots v_J) \\ &\quad \times \exp\left[-1/2(\ln(\frac{v_1 \sim v_2 \sim \ldots \sim v_{J-1}}{v_J} - \mu)'\Sigma^{-1}\ln(\frac{v_1 \sim v_2 \sim \ldots \sim v_{J-1}}{v_J} - \mu)\right] \end{aligned} \tag{13}$$

(where a $\sim$ denotes a column-wise concatenation of the v_j). Since the μ are unbounded, a reparameterization in terms of a linear combination of regressors is possible

$$\mu_j = X\beta_j, \tag{14}$$

where the X may again either be a set of identical regressors for each log-ratio (subject to identication constraints on Σ), or a different set for each log-ratio. If one assumes that the observations are identically and independently distributed, then the log-likelihood is

$$\begin{aligned} L(\beta,\Sigma|X,v) &= -3/2\ln(2\pi) - 1/2\ln(|\Sigma|) + \ln(v_1 v_2 \ldots v_J) \\ &\quad -1/2\left(\ln(v_1 \sim v_2 \sim \ldots \sim v_{J-1}/v_J) - \mu\right)'\Sigma^{-1}\left(\ln(v_1 \sim v_2 \sim \ldots \sim v_{J-1}/v_J) - \mu\right) \end{aligned} \tag{15}$$

This is essentially a seemingly unrelated regression model (SURE), with the addition of the jacobian $(\mathrm{jac}(v|y) = (y_1 \cdots y_{J-1})^{-1}$. (This means that conventional SURE methods can be used to obtain exploratory analyses, although standard error calculations will be incorrect without inclusion of the jacobian).

The advantages of this model, according to Aitchison, stem chiefly from the unconstrained properties of the covariance structure of the log-ratio transformed v. The new model now permits any pattern of dependence/independence among the time components y.

Our experience with estimation of this model has led us to discover that there are two non-trivial hindrances to application of this method. The first is that the method appears to be grossly sensitive to identification concerns. While the Dirichlet achieves identification through a strict pattern of covariation relationships and the form of the link functions, the generalized additive logistic normal must achieve identification through either exclusions in the log-ratio equations ($X\beta_j$), or (paradoxically) through restrictions on the covariance matrix (Σ), the very condition that induced us to consider something other than the Dirichlet.

The second is that there is no particular reason to assume that the log-ratio transformed compositions (the v) should be distributed normally. Whereas the Dirichlet distribution proceeds from first principles of independent gamma processes, it is not clear what first principals lead to a normal distribution of log-ratio compositions. Indeed, Katz and King (1997) in their application of these methods to analysis of multi-party elections in Britain conclude that the distribution of election results more closely patterns an additive logistic t distribution, although that is also not derived from first principles. Barring development of appropriate distributions based on first principles, the implementation of the additive logistic method requires an ad hoc search for distributions based on fit.

3 Simulation Comparisons

Our approach in this paper is to use Monte Carlo simulations to gauge when application of either Dirichlet or additive logistic methods would be inappropriate (i.e., result in misleading estimates for the overall composites, or for the effect of covariates upon composites).

Why use Monte Carlo simulation methods to compare the two approaches for compositional data analysis? For certain purposes — namely, to assess bias, consistency, or efficiency of the estimators, or the sensitivity of those estimators to assumptions — it is *always* preferable to deductively produce closed form calculations of these properties, over simulation techniques.

But it is not always possible to produce the closed form calculations. Mathematical abilities constitute one such limit (and the limits of the present authors preclude closed solutions to a direct comparision). Some problems, too, are precluded from deductive solutions because closed form solutions do not exist. In the present case, we note that the first derivative of the gamma function is represented in most statistics texts as a psi function without closed representation.

But there are more positive beneifts to simulation approaches. Provided one constructs reasonable facsim-

iles to real world problems, simulation approaches can yield more concretely comprehensible representations of the boundaries of appropriate analysis than deductive approaches. Aitchison's proofs, for example, hinge on probability limit calculations. We have 1000 observations — are we close enough to the asymptote? at this many observations, how much bias would we encounter? when are we most in jeopardy by the strong IIA assumptions?

The keys for us are a) to devise facsimiles of plausible settings for compositional data analysis problems, b) not to be satisfied with results that "work" (i.e., appear to be unbiased, with low error dispersion), but to push the scenarios until we find those that "fail" (i.e., appear to be biased, to be sensitive to assumptions, to leave unacceptably wide error dispersion).

We will generate composite distributions (for three composites, an obvious extension of the analysis is to greater numbers of composites) based on the following variations:

- Covariates in common across all composites ($\mu_i = \exp(X\beta_i)$);
- Covariates unique to composites ($\mu_i = \exp(X_i\beta_i)$).

Generation of composites requires attention to the covariance between log-ratios (and, hence, components y_j). Begin by computing the mean of each log-ratio:

$$\mu_j^* = \hat{\beta}_{0j} + \hat{\beta}_{1j}\bar{x}_{1j} + \ldots + \hat{\beta}_{kj}\bar{x}_{kj}. \tag{16}$$

Generate U as a matrix of $J-1$ normally distributed vectors of 1000 observations, each with mean μ_j^* and variance of 1. Each of the columns of U is independent, but with correct mean and variance for the log-ratios. If one multiplies U by the Cholesky decomposition of Σ, one can generate the correct log-ratios:

$$V = U\text{chol}(\Sigma)'. \tag{17}$$

We then convert back from the V to y_j by application of equation 12. Because Σ must be positive definite for the Cholesky decomposition to exist, we are constrained in our choice of appropriate matrices. We fix the main diagonal to 1, and the off-diagonal to the range $-1 \ldots +1$. We begin the analysis to three composites, although an obvious extension is to more than three, the choice of Σ becomes more complex.

The first simulation begins by first drawing three regression parameters $(\beta_0, \beta_1, \beta_2)$ from a uniform distribution (0,2), and a pair of covariates (x_1, x_2) from a uniform distribution (0,1). We then generate the μ_i by taking $\exp(\beta_0 + \beta_1 x_1 \beta_2 x_2)$. We then draw 1000 observations from a normal distribution with

mean μ_i. We then multiply the log-ratios by the Cholesky decomposition of Σ, drawn as described in the preceeding paragraph. We then transform the now correlated log-ratios back to composites. This method yields composites which vary from .2 to .8, with covariation between composites which varies from $-.8$ to .5.

The second simulation draws four regression parameters and three covariates from uniform distributions (with the same bounds as the second simulation). We then generate the ν_i by taking $\exp(\beta_0 + \beta_i x_i)$, that is, allowing only one x to be associated with each composite.

For both of the simulations, we then estimate with both Dirichlet and additive logistic normal, repeating for 1000 replications.

The coefficients across the Dirichlet and additive logistic normal methods are not directly comparable. Not only are these on different scales of different functional forms, but we have three sets of coefficients for the Dirichlet and two for the additive logistic normal. Instead, we estimate first differences, computed as the change in the composite for each of the x_i, with x_i varying from the minimum (0) to maximum (1), holding all the other x at the mean (.5).

4 Results

4.1 Common Covariates

In the first simulation, we let the parameters for the compositions be a function of three randomly selected parameters (a constant plus two slope terms) and randomly selected X, all drawn from a uniform distribution, with the X in common across all three composites. The idea in this simulation is to apply both Dirichlet and additive logistic methods to estimate the effect of a unit change in each of the two Xs on the composite against the true effect.

Figure 8 presents the kernel density plots for the error in estimates of effects for both Dirichlet and Additive Logistic methods (where the error is the true effect minus the estimated effect). The upper half of the plots contain the Dirichlet estimates, and are noted with a **(D)**; the lower half of the plots contain the additive logistic estimates, and are noted with a **(A)**. Plainly, the error in estimates of effects is quite small, on average about zero, approximately normally distributed and approximately the same for both Dirichlet and Additive Logistic approaches.

In figure 9, we plot the estimated effect against the true effect, with a line drawn at 45 degrees to denote equality. As is implicit from the previous graph, the points are clustered about the 45 degree line. There is

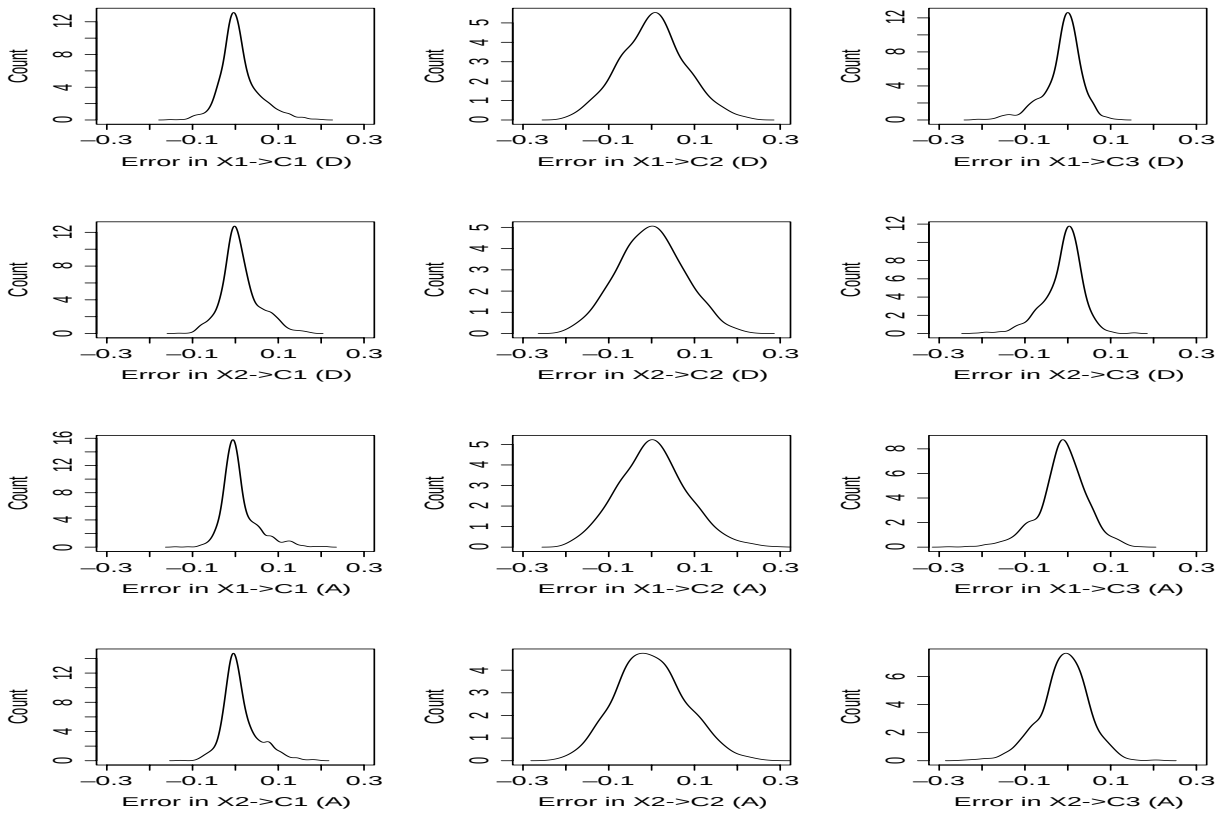

Figure 8: Error in Estimates of Effects on Composites, Common Covariates Simulation (Kernel Density Plots)

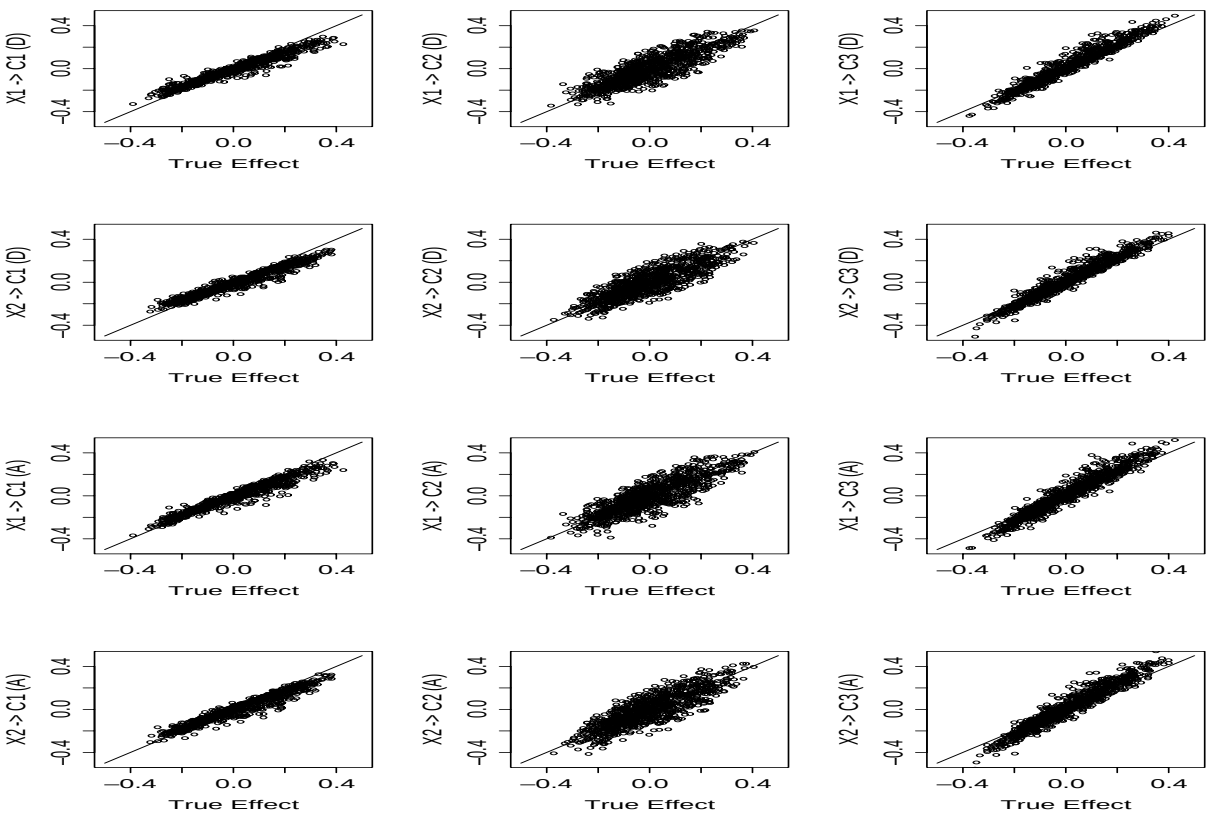

Figure 9: Estimate vs. True Effect on Composites, Common Covariates Simulation

slight evidence of bias to the Dirichlet estimates, in that the cluster veers from the 45 degree line as the true effect veers toward the extremes.

We can directly identify what accounts for greater divergence from the true effects. Figure 10 presents the error in estimated effects plotted against the correlation between the log-ratios. As the log-ratios become more strongly correlated, the dispersion of the error in estimated effects increases. But the problem holds for *both* the additive logistic and Dirichlet approaches, and appears to be as likely to be an overestimate as an underestimate of effects.

In brief, then, when the composites are influenced by common covariates, there appears to be no advantage in use of the additive logistic methods over the Dirichlet.

4.2 Unique Covariates

The second simulation employed in this paper is to assign composites by unique covariates. We draw two random coefficients (a constant plus one slope term) and three random Xs. The parameters for the

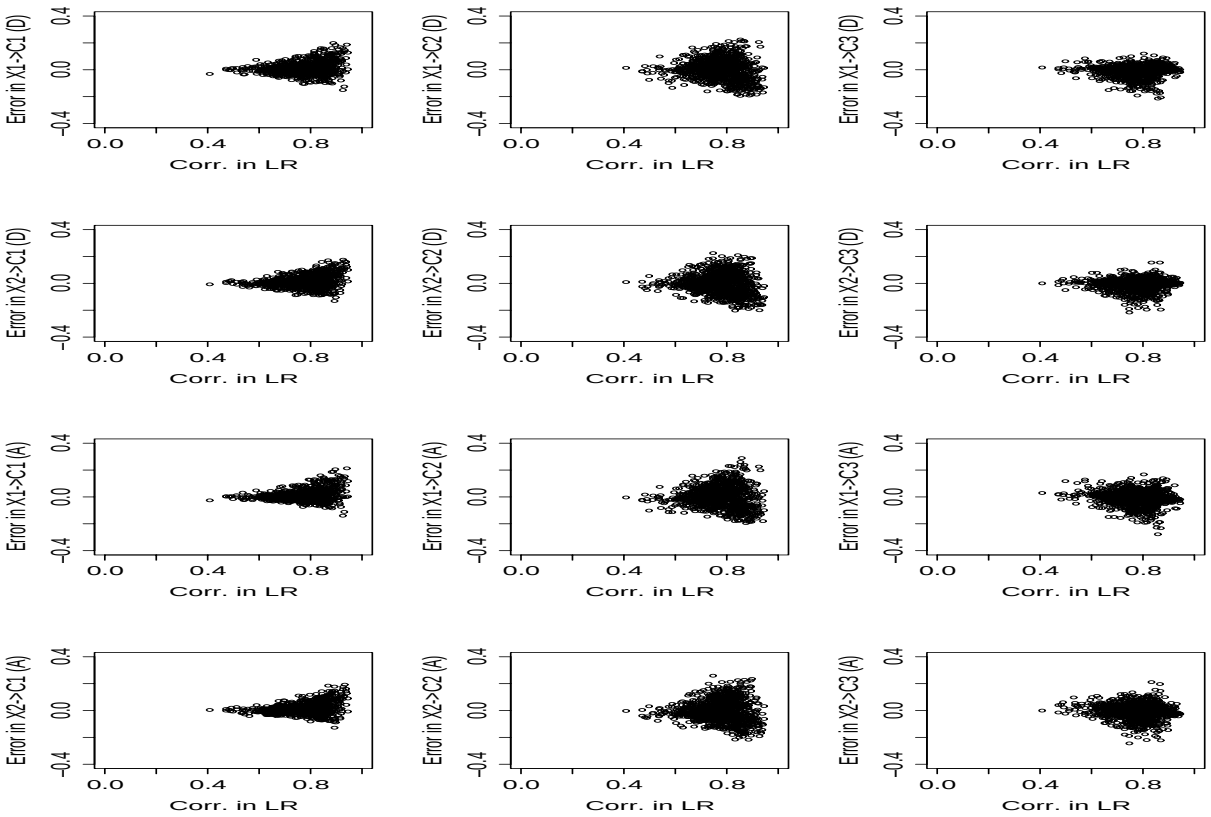

Figure 10: Error in Estimates of Effects on Composites vs. Covariance of Log-Ratio of Composites, Common Covariates Simulation

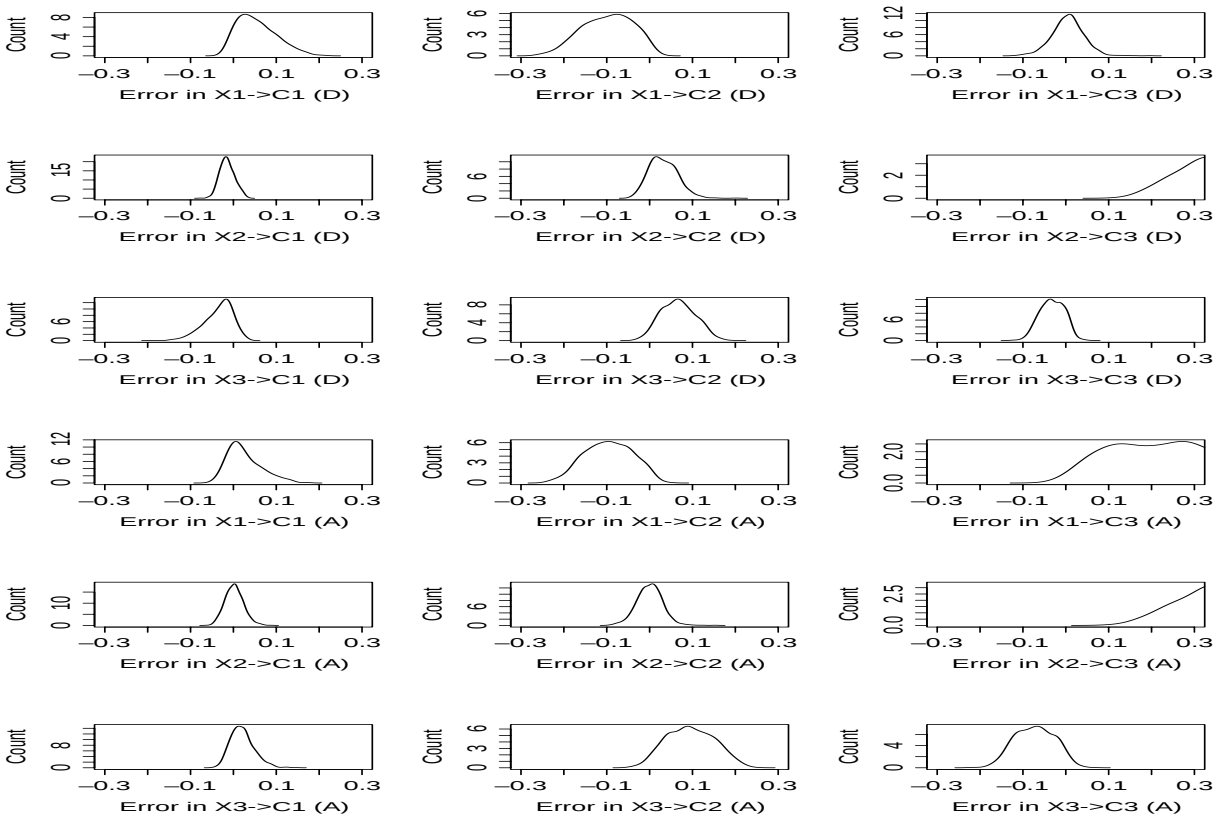

Figure 11: Error in Estimates of Effects on Composites, Unique Covariates Simulation (Kernel Density Plots)

random draws of composites are a function of an X unique to that parameter (i.e., $X1$ for composite 1, $X2$ for composite 2, and $X3$ for composite 3). The intuition to be tested here is that any problems in mis-estimating effects should be most obvious when composites are highly correlated.

Turn first to the kernel density plots of the error in estimates of the first differences for both the Dirichlet and additive logistic estimates (figure 11). As before, the Dirichlet estimates are in the upper half of the figure and marked with a **(D)**, while the additive logistic estimates are in the lower half of the figure and marked with an **(A).**

The intuition we have is that the estimates for the effect of covariates that are associated with each composite should be reasonably accurate, but that the estimates for the effect of other covariates may be erroneous (and, further, more biased as the covariation between the composite in question and the composite uniquely associated with the covariate increases). In other words, we expect low variance on the $X1 \rightarrow C1, X2 \rightarrow C2$, and $X3 \rightarrow C3$ effect estimates (call these the "on target" covariates), and wider variance otherwise ("off target").

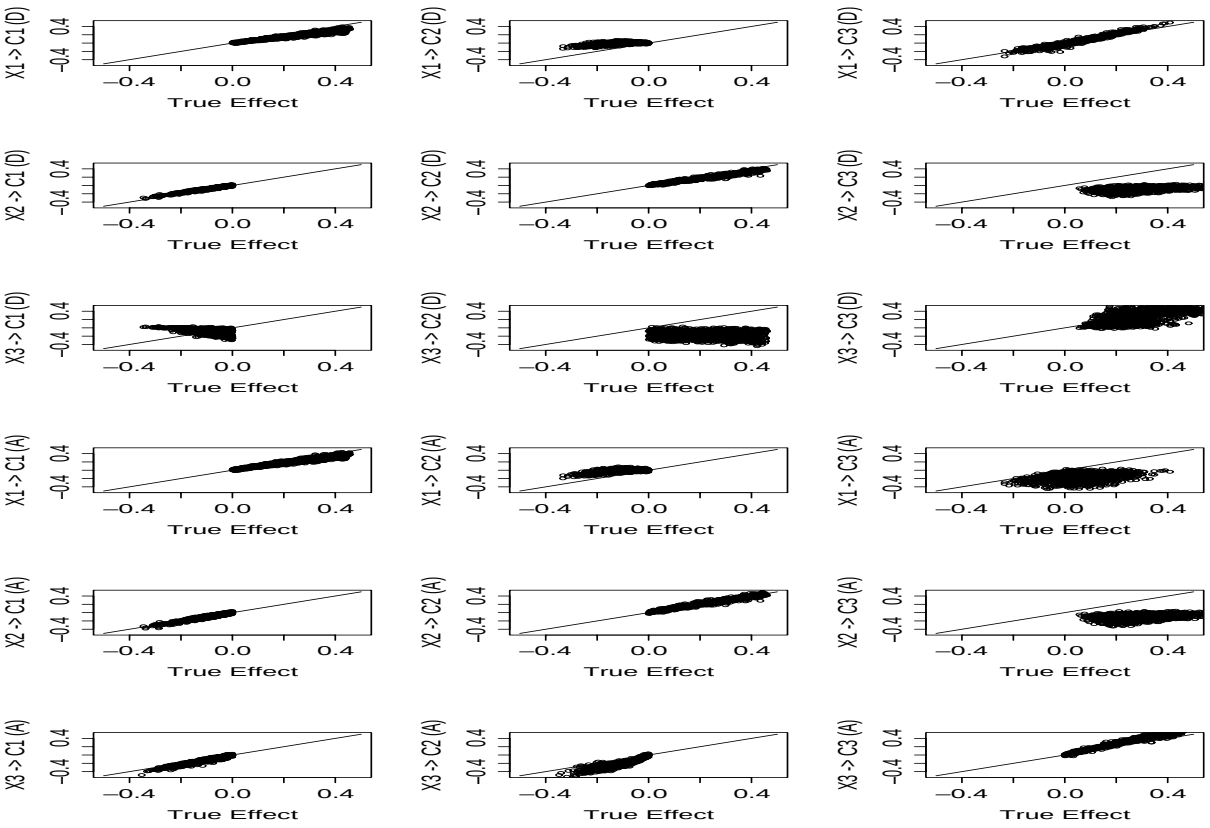

Figure 12: Estimate vs. True Effect on Composites, Unique Covariates Simulation

The actual results are somewhat surprising. The "on target" estimates are quite close to the true estimates: the mode is essentially zero, and the dispersion quite narrow. This holds for both the Dirichlet and additive logistic forms. The "off target" estimates are all over the map. The majority of the estimates exhibit very small levels of bias (noted by the difference of the mode from zero), with narrower error dispersion in the additive logistic form as in the Dirichlet. But there are three cases where the kernel density plots suggest severe problems: the Dirichlet estimates of the effects of $X2$ on $C3$, and the additive logistic estimates of the effects of $X1$ and $X2$ on $C3$. We do not have any particular reason to offer for the difficulties in estimating $C3$, although this is the composite which is determined by the other two for the additive logistic case (i.e., as $1/(1+\exp(X\beta_1)+\exp(X\beta_2))$).

The next set of plots provide more complete clues as to the nature of bias in the simulated estimates (figure 12). Again, we would like to see the estimated effects clustered close to the diagonal marking equality. Here, the Dirichlet approach fares substantially poorer than the additive logistic. The "on target" estimates for the Dirichlet are symmetrically arranged around the diagonal, although they are widely dispersed for the

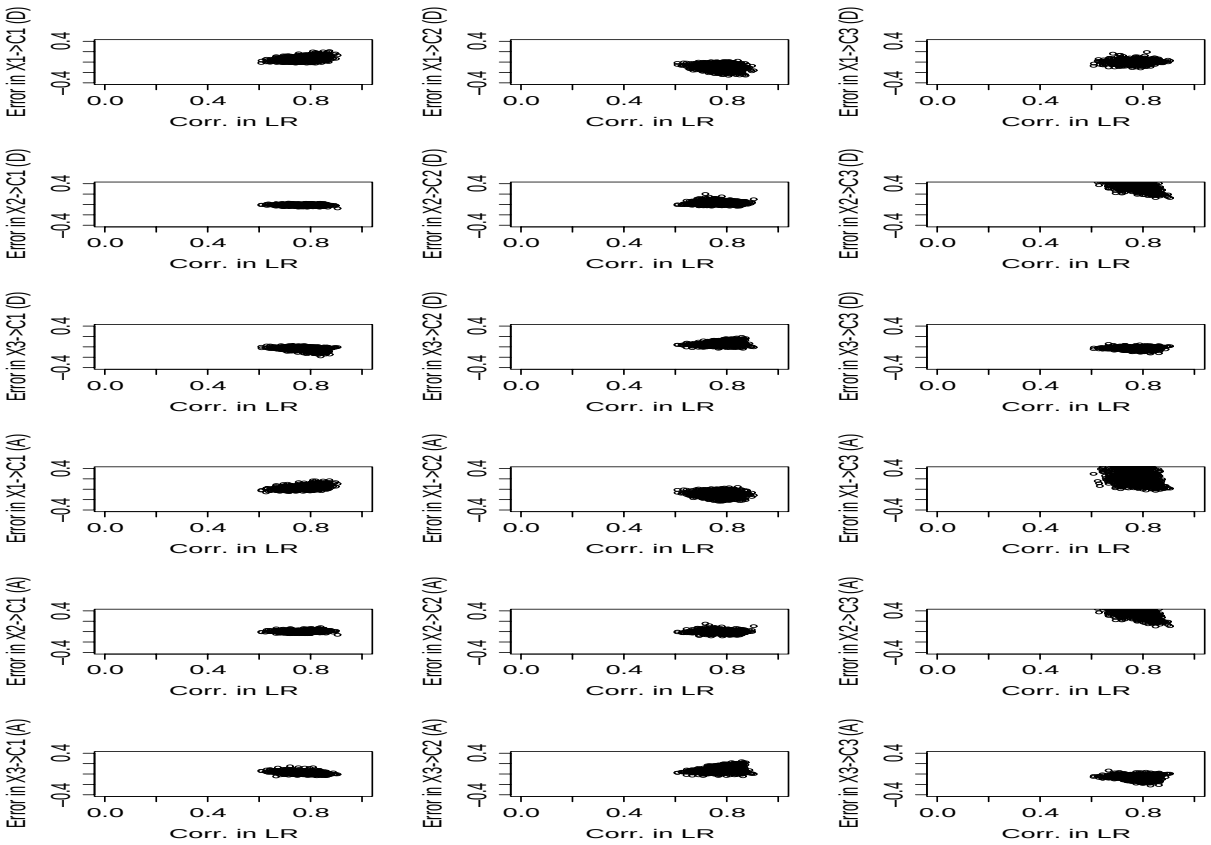

Figure 13: Error in Estimates of Effects on Composites vs. Covariance of Log-Ratio of Composites, Unique Covariates Simulation

effect of $X3 \rightarrow C3$. The key here is that only two of the "off target" estimates exhibit the same pattern of symmetrical arrangement around the diagonal ($X1 \rightarrow C3, X2 \rightarrow C1$), while the remainder are quite off the mark.

The additive logistic estimates, by contrast, appear to be much closer to the mark. The "on target" estimates are all close to the diagonal, symmetrically distributed. Only two of the "off target" estimates are seriously off the mark.

The final set of plots (figure 13) explore the relationship between the error in estimates for both methods as a function of the correlation between the log-ratios. As before, ideally, the graph should be flat and fairly evenly concentrated around zero. When the dispersion increases as the correlation increases, this is evidence of increased sensitivity to the correlation between log-ratios of composites.

Lke the equivalent graphs for the common covariates simulation, the plots of error against the correlation in log-ratios imply increasing sensitivity to the correlation between log-ratios. All but one of the Dirichlet

estimates demonstrate general sensitivity to the correlation in log-ratios. The remaining plot (for $X2 \rightarrow C3$) is far off the mark, with a frankly odd pattern. And again, the additive logistic estimates appear to be *equally* sensitive to the correlation in log-ratios. One of the estimates ($X2 \rightarrow C3$) is extremely off the mark, with another frankly bizarre pattern.

Are the odd patterns (for both the Dirichlet and additive logistic effects) more comprehensible when framed as errors against the correlations of the composites themselves? Nope. We have explored the plots of the error against the correlation between the composites themselves. There is, however, only a weak relationship between the correlations of composites and effects seen. As the covariation for composites 1 and 2 increases, the estimate of the effect of $X1 \rightarrow C3$ becomes more biased, but no others, for both Dirichlet and additive logistic approaches. As the covariation between composites 1 and 3 increases, the Dirichlet estimates of $X1 \rightarrow C3$ and the additive logistic estimates of $X1, X2 \rightarrow C3$ become more biased, but no others.

In short, the additive logistic and Dirichlet methods appear to be approximately equally successful at estimating the effects of the unique covariates on the composites: they are generally on target, with low, and symmetric error distributions. There are a few odd cases where estimates fall far from the mark, but there is no obvious advantage to either method. More so, both methods become more error-prone as the correlation between the log-ratios increases — exactly the evidence of the IIA assumption. Although the additive logistic method purports to handle strong conditions of dependence between ratios of composites, in these simulations, it fairs no better than the Dirichlet.

5 Conclusion

So, what to make of the comparisons between the methods? Recall the goal of our simulations: to devise situations in which the Dirichlet method should be likely to fail, while the additive logistic method should be likely to succeed.

We failed at failing.

Unfortunately, with simulation results, that is not the same as succeeding. Perhaps we did not devise the right scenarios to demonstrate the failings of Dirichlet methods, or the strengths of additive logistic. Perhaps with a greater number of composites, a more complex pattern of intercorrelations among composites, or nonlinear functional forms for the effects of variables upon composites we would see more stark evidence of

the strengths of one approach versus the other.

The problemof biased estimates surfaces most obviously in the condition where each composite is determined by unique covariates. In our present substantive interest, we should be wary of interpreting effects for conditions where time to particular tasks (e.g., meals, paperwork) is determined by highly task-specific variables.

We do, however, see clear strengths for the Dirichlet approach within the quite narrow bounds of these simulations. The Dirichlet approach is significantly faster than additive logistic. Total estimation per replication (on average, on a Pentium Pro 180 running Gauss under Linux) required 60.7 seconds for the Dirichlet estimates and 221.3 for the additive logistic estimates for the unique covariates simulation, for these extremely simple models. Our experience in estimating the two sets of models for the police data is that these differences add up to hours of advantages in favor of the Dirichlet. Hours of improved speed mean more model testing, more model variation, more specification approaches, and thus better science.

The Dirichlet distribution is considerably more flexible than the additive logistic. One method of demonstrating results is to produce simulations of plausible values. The additive logistic estimates, in our experience with the police data, produce overestimates of the dispersion of police performance — substantively quite important, since convergence among officers is partial evidence for intersubordinate influence. The Dirichlet estimates, on the other hand, generated much more accurate pictures of their performance (see figures 14 and 15).

These are simulation results, and provisional evidence. Others (Rayens and Srinivasan 1994) claim that another distribution (the Liouville) can capture the extreme virtues of both methods. (It is, however, subject to some dispute as to whether the density integrates within legitimate probability bounds). Our evidence suggests that there is no particular advantage to additive logistic methods over Dirichlet. Our experience suggests that the methods yield rich opportunities for research into interesting political problems.

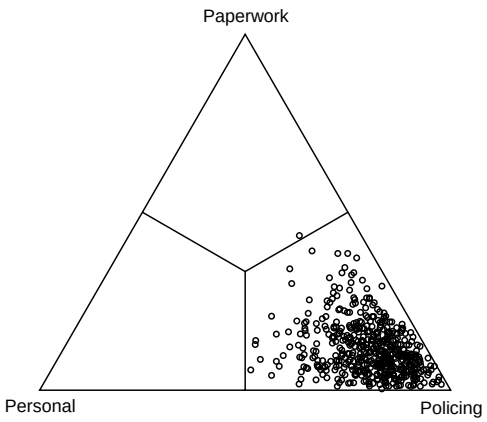

Figure 14: Simulated Distribution of Police Time at Mean Dirichlet Estimates

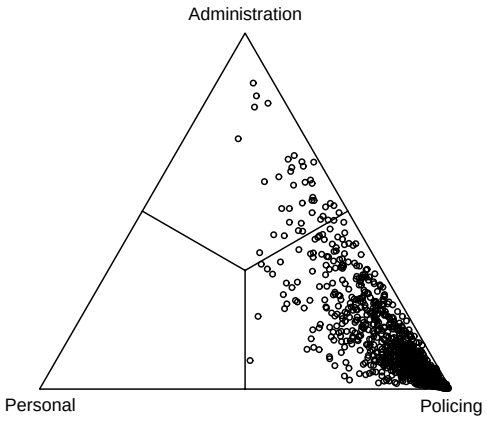

Figure 15: Simulated Distribution of Police Time at Mean Additive Logistic Normal Estimates

6 References

Aitchison, J. 1986. *The Statistical Analysis of Compositional Data* New York: John Wiley.

Aitchison, J. and S. M. Shen. 1980. "Logistic-normal Distributions: Some Properties and Uses," *Biometrika* 67: 261-72.

Brehm, John, Scott Gates, and Brad Gomez. 1998. "Donut Shops, Speed Traps, and Paperwork: Supervision and the Allocation of Time to Bureaucratic Tasks." Paper presented at the 1998 Public Choice Society Annual Meeting and the 1998 Midwest Political Science Association Annual Meeting.

Brehm, John and Scott Gates. 1997. *Working, Shirking, and Sabotage: Bureaucratic Response to a Democratic Public*. Ann Arbor: University of Michigan Press.

Gupta, Rameshwar D. and Donald St. P. Richards. 1987. "Multivariate Liouville Distributions," *Journal of Multivariate Analysis* 23: 233-56.

Katz, Jonathan and Gary King. 1997. "A Statistical Model of Multiparty Electoral Data," Paper presented at the annual meetings of the Midwest Political Science Association, Chicago, April.

Ostrom, Elinor, Roger B. Parks, and Gordon Whittaker. 1988. *Police Services Study, Phase II, 1977: Rochester, St. Louis, and St. Petersburg*. ICPSR 8605.

Padgett, John F. 1981. "Hierarchy and Ecological Control in Federal Budgetary Decision Making." *American Journal of Sociology* 87(1): 75–129.

Rayens, William S. and Cidambi Srinivasan. 1994. "Dependence Properties of Generalized Liouville Distributions on the Simplex," *Journal of the American Statistical Association* 89: 1465-70.

Sivazlian, B. D. 1981. "A Class of Multivariate Distributions," *Australian Journal of Statistics* 23, 2: 251-5.