Published in final edited form as:

Comput Stat Data Anal. 2008 August 15; 52(12): 5344–5355. doi:10.1016/j.csda.2008.05.030.

Dirichlet Component Regression and its Applications to Psychiatric Data

Ralitz Gueorguieva¹, Robert Rosenheck², and Daniel Zelterman^{1,*}

¹Division of Biostatistics, Department of Epidemiology and Public Health Yale University, New Haven, CT 06520

²VA Connecticut Healthcare System, West Haven, CT

Summary

We describe a Dirichlet multivariable regression method useful for modeling data representing components as a percentage of a total. This model is motivated by the unmet need in psychiatry and other areas to simultaneously assess the effects of covariates on the relative contributions of different components of a measure. The model is illustrated using the Positive and Negative Syndrome Scale (PANSS) for assessment of schizophrenia symptoms which, like many other metrics in psychiatry, is composed of a sum of scores on several components, each in turn, made up of sums of evaluations on several questions. We simultaneously examine the effects of baseline socio-demographic and co-morbid correlates on all of the components of the total PANSS score of patients from a schizophrenia clinical trial and identify variables associated with increasing or decreasing relative contributions of each component. Several definitions of residuals are provided. Diagnostics include measures of overdispersion, Cook's distance, and a local jackknife influence metric.

Keywords

Multivariable regression; overdispersion; Cook's distance; local influence; PANSS

1 Introduction

Much of the quantitative aspects of psychiatry consists of a trained professional asking the subject a series of questions, and grading each on a scale of 1 to 5 or 1 to 7. The separate grades are then added together giving a single numerical composite score of mental well-being or illness severity. While changes in the sum will quantify global improvement or deterioration, it is often of interest to examine different aspects of the global change and to simultaneously assess the relative contribution of each aspect as a fraction of the total. As an example, symptoms and behaviors characteristic of schizophrenia can be assessed using the Positive and Negative Syndrome Scale (PANSS) (Kay *et al.*, 1989). The sum of all individual question scores is the total PANSS which is usually broken down into subscales that define different aspects of overall disease severity. There is often interest in identifying the association of baseline socio-demographic and co-morbid correlates of measures such as

© 2008 Elsevier B.V. All rights reserved.

*Corresponding author: daniel.zelterman@yale.edu, tel: +1 203 785 5574, fax: +1 203 785 6912.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

the total PANSS score, to identify factors related to increased disease severity. In such analyses multiple multivariate models are often evaluated to identify correlates of the total score and each of the subscales. However, methods have yet to be demonstrated that can identify in a single parsimonious model the association of baseline correlates with increasing or decreasing proportions of the total severity score that are attributable to each of the component subscales. This study demonstrates the use of Dirichlet models of multivariable linear regression that can address this more complex set of relationships.

The data that motivated our analysis were part of a randomized, one-year, double-blind comparative study of clozapine and haloperidol at 15 Veterans Affairs medical centers (Rosenheck *et al.*, 1997). For simplicity we focus only on the baseline data. All participants were previously diagnosed with refractory schizophrenia and their symptom severity was measured by the PANSS.

The total PANSS score consists of the sum of scores on each of 30 questions, each evaluated on a scale of 1 to 7 with higher numbers always representing greater disease severity. We analyzed PANSS data according to a previously derived, empirical PANSS factor structure grouping the questions into five non-overlapping component scales described as positive, negative, cognitive, emotional and hostility factors (Bell *et al.*, 1994). One of the questions (disorientation) does not load substantially on any of the five factors and hence is excluded from the analysis.

The association with variable labels and marginal summary statistics of the five components are given in Table 1. Each of the five components X_i is composed of the sum of between 4 and 8 questions and is measured on a different scale. On every component as well as their sum, lower values represent lower symptom severity. In the present work we will treat these individual component values as continuous. The marginal distributions of these five components are plotted in Fig. 1.

The following baseline explanatory variables were available to us: age, gender, race, duration of illness, lifetime history of depression, lifetime history of substance abuse, Simpson-Angus score and AIMS score. The 423 subjects were 98% male and the small numbers of female subjects were later dropped from our analysis. A large percent of this patient population had a lifetime history of substance abuse and 7% recorded a lifetime history of ever having been diagnosed with depression. Ages ranged from 23 to 76 years with a mean of 44 and a standard deviation of 8 years. Race was binary-coded as white and non-white. The duration of illness data contains missing values on 13% of the subjects but, as we see in Section 6.1, this variable provides a large amount of explanatory value so we keep it under consideration. The Abnormal Involuntary Movement Scale (AIMS) is a measure of involuntary movements related to Tardive Dyskinesia (*e.g.* writhing movements of the limbs or protruding movements of the tongue). Such movements sometimes develop as a side effect of long-term treatment with antipsychotic medications. The Simpson-Angus instrument measures neuroleptic-induced parkinsonism (*e.g.* stiffness and tremors, Simpson, Angus, 1970). This measure has a skewed distribution and we chose to work with the log transformation of the original values.

In this work we seek to assess the effects of baseline covariates on the mean and variability of the five individual component subscales of the PANSS as fractions of the total PANSS. In particular, if $\{X_1, \dots, X_5\}$ represent the five individual components then

$$\text{PANSS} = \sum_{i=1}^5 X_i$$

is defined as the PANSS total score.

We jointly model the fractional components

$$Y_i = X_i / \text{PANSS}$$

using a Dirichlet distribution model and linear predictors of baseline covariate variables. This component analysis of $\{Y_1, \dots, Y_5\}$ is secondary to the examination of the total PANSS. Since each of the subscales consists of several individual questions, a hierarchical Dirichlet regression is also possible.

Dirichlet regression models and estimation for such models are considered by Campbell and Mosimann (1987) and Hijazi (2003) and can be regarded as generalization of beta regression models (Ferrari and Cribari-Neto, 2004) for more than two components. Dirichlet regression is particularly suited for the analysis of compositional data and is an alternative to the approaches of Aitchison (1986) and Barcelo *et al.* (1996) for such data. In these latter approaches either additive normality log-ratio or Box-Cox transformations are used so that traditional multivariate techniques can be applied to the transformed data for estimation and inference. Barcelo *et al.* (1996) point out that the choice of the transformation is not straightforward and can be influenced by outliers.

For Dirichlet regression models, Campbell and Mosimann (1987) and Hijazi (2006) propose model-fit and diagnostics measures. Hijazi (2006) proposes to use quantile residuals to assess distributional assumptions and considers three R^2 type measures of total variability attributable to the model predictors and two measures of goodness-of-fit. He also suggests a likelihood distance measure to identify influential compositions of observations. For beta regression models, there are a number of recently proposed residuals and influence measures. Ferrari and Cribari-Neto (2004) develop standardized residuals and residual plots, a global measure of discrepancy-of-fit and local influence measures. Espinheira *et al.* (2008a) propose two new residuals, numerically evaluate them and demonstrate the advantages of the standardized weighted residual. Espinheira *et al.* (2008b) develop several Cook-like measures for assessing influence under different perturbation schemes.

There is little overlap between the diagnostics and graphical measures that we propose for Dirichlet regression models and the work of Hijazi. We focus on identifying outlying/influential observations at the individual data level through standardized and score residuals, and using star and tree plots. We also present local influence measures and focus on overdispersion in assessing overall model fit.

Our manuscript extends and presents alternatives to some of the diagnostics developed by Ferrari and Cribari-Neto (2004) and Espinheira *et al.* (2008ab) for the beta distribution. Our standardized residuals are generalizations of the standardized residuals of Ferrari and Cribari-Neto. Like the weighted residuals of Espinheira *et al.* (2008a) our score residuals are obtained in the maximum likelihood estimation process. We also propose two multivariate residuals. In developing our local influence measures we focus on assessing influence due to deleting or down-weighting a single observation which is one of the scenarios considered by Espinheira *et al.* (2008b). It is possible to extend our work to generalizations of the local influence measures proposed by Espinheira *et al.* (2008b) for beta regression under alternative perturbation schemes. In the beta regression standard residual plots may be used to assess deviations from assumptions. In contrast, the multivariate nature of the residuals from Dirichlet models leads to the development of unique residuals plots such as the star, spider and tree plots proposed here.

Alternative models to the Dirichlet for compositional data are explored by Aitchison and Shen (1980) and Shen (1980). Both of these methods propose using a multivariate normal distribution with a logistic link function for the mean and an unstructured variance matrix. Such an approach would have too many parameters for the data considered here, where every individual has a unique set of covariate values. In contrast, all of the numerical examples in Aitchison and Shen (1980) involve small numbers of different covariate values. The generalized Dirichlet model proposed by Connor and Mosimann (1969) is only useful if the fractional components Y_i have an implicit ordering.

The Dirichlet model is introduced in Section 2, including the likelihood and several definitions of residuals. Measures of influence and overdispersion are described in Section 5. We return to the original example in Section 6.

2 Notation and Methods

The Dirichlet distribution is often used as a prior in the Bayesian analysis of multinomial data and not often thought of as a response distribution. The Dirichlet distribution is a generalization of the beta distribution to higher dimensions. Let $\{Y_1, \dots, Y_k\}$ denote a random k -tuple ($k \geq 2$) with all $0 \leq Y_i \leq 1$ and $\sum Y_i = 1$ for $i = 1, \dots, k$.

The multivariate density function of the Dirichlet distribution is

$$f(y_1, \dots, y_k | \alpha_1, \dots, \alpha_k) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} \prod_i y_i^{\alpha_i - 1} \quad (2.1)$$

for parameters $\alpha_i > 0$ defined on $0 \leq y_i \leq 1$ with $\sum y_i = 1$.

When $k = 2$ expression (2.1) reduces to a beta distribution. The marginal distribution of a single Y_i in this distribution also has a beta distribution. When all α_i are proportionately large then (2.1) can be approximated by a multivariate normal density.

The expected value of each Y_i is

$$EY_i = \alpha_i / \sum_j \alpha_j$$

and the variance is

$$\text{Var}Y_i = EY_i(1 - EY_i) / \left\{ 1 + \sum_j \alpha_j \right\}. \quad (2.2)$$

Jointly, the $\{Y_i\}$ is constrained to sum to one so that the covariance between any pair

$$\text{Cov}(Y_i, Y_{i'}) = -E(Y_i)E(Y_{i'}) / \left\{ 1 + \sum_j \alpha_j \right\} \quad (2.3)$$

with $i \neq i'$ is always negative.

Generally speaking, we see that the variance of Y_i is larger when its mean is closer to 1/2, as is also the case with the beta distribution. The variances of the $\{Y_i\}$ are smaller for larger value of $\sum \alpha_i$. As with many other distributions, the parameters that describe the mean are also related to the expressions for the variance. It is possible to find that in modeling, the $\{\alpha_i\}$ parameters can all increase but the means of the $\{Y_i\}$ are relatively unchanged or perhaps some mean responses may decrease. An example of this phenomenon is illustrated in Fig. 2. Intuitively, the means of the $\{Y_i\}$ must sum to one so not all can simultaneously increase or decrease.

The models we develop for the fractional components $\{Y_i\}$ of PANSS describe each $\log(\alpha_i)$ as a separate linear function of covariates (such as age and duration of illness of each subject) and regression coefficients. That is, for each component $i = 1, \dots, k$ we use a log-link with

$$\log \alpha_{ij} = \beta_i' z_j \quad (2.4)$$

for covariates z_j recorded on the j -th individual ($j = 1, \dots, n$) and regression coefficients $\{\beta_i\}$ to be estimated using maximum likelihood.

Model (2.4) will usually contain an intercept for each α_i . It is possible to include covariates that are specific to each of the component values as later illustrated for our example in Table 2.

The β_i regression parameters can be fitted using maximum likelihood with estimates denoted $\widehat{\beta}_i$ and corresponding estimates $\widehat{\alpha}_i = \{\widehat{\alpha}_{ij}\}$ of $\alpha_i = \{\alpha_{ij}\}$ defined by

$$\widehat{\alpha}_{ij} = \exp(\widehat{\beta}_i' z_j)$$

making use of the inverse of the log-link function (2.4).

We will refer to the $\{\alpha_i\}$ as *meta-parameters* because these combine the effects of the covariates z_j using regression parameters $\{\beta_i\}$ that are estimated by the software in the application. We performed most of the computing in SAS using the general likelihood model statement in the nlmixed procedure to obtain maximum likelihood estimates. We did not explore properties of models with random effects such as might be used to model the longitudinal aspects of PANSS following medical intervention over time.

We next describe several different definitions of residuals for these models: three univariate or marginal residuals and two multivariate residuals. The simplest form of univariate residual is the *raw residual* or difference d_{ij} between observed and expected

$$d_{ij} = y_{ij} - E(Y_{ij} | \widehat{\alpha}_j).$$

These residuals sum to zero over components i within every individual j but not necessarily across individual subjects. The raw residuals are difficult to compare because they lack a comparable variance. It is better to compare the values of *standardized residuals* defined as

$$r_{ij} = (y_{ij} - E(Y_{ij} | \widehat{\alpha}_j)) / \sqrt{\text{Var}(Y_{ij} | \widehat{\alpha}_j)}.$$

The normalizing variance of the marginal standardized residuals does not take into account the reduced variability associated with using estimated values instead of the true values α but these remain our preferred method of identifying univariate outliers.

The *composite residual* combines standardized residuals within the j -th subject

$$C_j = \sum_i r_{ij}^2$$

much as we would compute the Pearson chi-squared statistic.

The composite C_j gives an equal weight to the different within-subject standardized residuals. Cook (1986) suggests that we combine residuals or influence measures weighted by eigenvectors of their correlation matrix. Such a weighting will allow for detection of *masking*, a setting in which large outliers may be hidden behind more moderate outliers. In the data analysis of Section 6 and in Fig. 6 we will demonstrate that C_j has a high correlation with Cook's distance in these data.

The *Mahalanobis* composite residual M_j is a weighted distance (see, for example, Rao, 1973, p 595, or Baxter, 1999) between the raw residuals $d_j = \{d_{ij}\}$ on the j -th subject and the origin. It is defined by

$$M_j = d_j' \widehat{\Sigma}_j^{-1} d_j$$

where $\widehat{\Sigma}_j^{-1}$ is a generalized inverse of the estimated variance of d_j defined at (2.2) and (2.3) using the maximum likelihood estimates of α_j . The inverse of the singular matrix $\widehat{\Sigma}_j$ is not defined because the raw residuals d_j sum to zero within every subject.

In addition to the intuitive, standardized residuals that are based on differences between observed and expected values there are also the score residuals that are obtained as part of the maximum likelihood process. The *score residual* is defined as

$$s_{ij} = \psi \left(\sum_i \widehat{\alpha}_{ij} \right) - \psi \left(\widehat{\alpha}_{ij} \right) + \log y_{ij} \quad (2.5)$$

where $\psi(x) = \partial \log \Gamma(x) / \partial x$ is also known as the *digamma function* (Abramowitz and Stegun, 1972, Section 6.3).

The score residuals sum to zero over subjects (in j) for each component i as a consequence of the maximization of the likelihood in α . The score residuals do not sum to zero within a subject as is the case with the raw residuals. The details and derivation of the score residuals are given in the Appendix. In general, we found that there was extremely high correlation between the corresponding standardized r_{ij} and score residuals s_{ij} . This correlation was always greater than .97 within each of the five components. The score residuals are also related to the leverage measures described in Section 5.

3 Graphical Representations of Raw Data

The mosaic plot was developed by M. Friendly and is useful for depicting the component data described here. These methods are reviewed and generalized in Friendly (1999) for categorical data.

Fig. 3 depicts a mosaic plot for seven individuals with the largest values of one of the covariates (the Simpson-Angus). Each individual's responses on the five components is plotted as a column of five tiles or mosaics. The width of the mosaics within each column is constant. Each individual's total PANSS score is given at the bottom of the figure. The height of each mosaic is proportional to the amount each component question contributes to the total PANSS score. The indexes of the questions in Table 1 are listed along the left edge of this figure and are ordered by the proportional contribution of Y_1 : PANSS Positive.

As an example of interpreting Fig. 3 consider the individual with a total PANSS of 126, depicted at the far right of this figure. In comparison with the six other patients whose data are plotted here, we see that this individual exhibits proportionately large values of Y_2 : Negative and Y_3 : Cognitive but relatively small proportions of Y_1 : Positive and Y_5 : Hostility. In contrast, the patient depicted at the far left Fig. 3 with a total PANSS of 83 exhibits a small proportion of Y_3 component but a very large Y_1 component.

4 Tree and Star Plots

Star- and tree-plots are useful graphical representations of residuals following a regression analysis. Both the star-plot and mosaic plots are available in R.

These are plots that simultaneously display all univariate standardized residuals $\{r_i\}$ for a specified individual. Variations on the star plots are also known as radar, spatial star, and spider plots.

Tree plots depict the separate residuals as branches growing off a tree. An example is given in Fig. 4 for the individual with the largest composite residual $C_j = 22.77$. Branches to the left are for negative residual values and positive valued residuals branch to the right. The vertical solid line gives the locations of zero residuals and the two dotted lines indicate the range for residuals equal to ± 2 . Dashed lines indicate the locations of residuals equal to ± 4 .

The data on the individual depicted in Fig. 4 shows that the standardized residual r_1 (PANSS Positive) has a huge negative residual (-3.13) and large positive standardized residuals for r_2 (2.73) and r_3 (1.71) and a moderately large negative residual for r_5 (-1.57).

The residuals for this same individual are displayed in the star plot of Fig. 5. The star plot in Fig. 5 depicts all of the residuals radiating out from the same point. The five directional axes are labeled. The dotted circles correspond to residual values of -2 for the innermost circle; 0 for the darker central circle; and $+2$ for the outermost circle. The values of the five adjacent numbered residuals are joined with straight lines producing this star-shaped figure. The star-plot given in Fig. 5 depicts the residual data for the same patient whose residuals are plotted in Fig. 4.

5 Diagnostics

We develop diagnostic metrics here to identify both overdispersion and high leverage in the data. The Appendix provides the mathematical details of the derivations, both for general likelihood-based inference, and with specific application to the Dirichlet component regression models described here.

5.1 Overdispersion

Overdispersion is an increase in the variance of the response variable anticipated by the model due to the lack of homogeneity in a specified parameter across subjects. Zelterman and Chen (1988) derived overdispersion statistics to test for parameter homogeneity when parameter values are allowed slight random variability between observations. While residual plots will often detect model deviations in terms of the means, these statistics are often useful for finding lack of fit involving larger than expected variances.

In the Dirichlet model without covariates the statistic for testing overdispersion of the **i-th Dirichlet component meta-parameter** α_i is

$$D_i = n\psi' \left(\sum \widehat{\alpha}_i \right) - n\psi' (\widehat{\alpha}_i) + \sum_{j=1}^n s_{ij}^2$$

where $\widehat{\alpha}_i$ is the maximum-likelihood estimate of α_i under the null hypothesis of equal α_{ij} parameters across subjects, s_{ij} is the score residual defined at (2.5), and $\psi'(x) = \partial^2 \log \Gamma(x) / \partial x^2$ is also known as the *trigamma function* (Abramowitz and Steegun, 1972, Section 6.4).

Each D_i is the difference between two measures of the observed information of α_i . Details of the derivation of D_i are given in the Appendix. Intuitively, D_i measures the difference between the observed curvature of the log-likelihood at $\widehat{\alpha}_i$ (as measured by the second derivative) and how large this curvature should be under the correct Dirichlet model. Larger values of D_i are indicative of less curvature, or a flattening of the log-likelihood at its maximum. The D_i statistics will generally have large values when the observed variances of the observations are larger than expected, perhaps explained by parameter values that vary slightly between observations. Large negative values of D_i are indicative of underdispersion and these are usually associated with model lack of fit elsewhere in the data. Under the null hypothesis of homogeneous values of $\{\alpha_i\}$, each statistic $D_i(\widehat{\alpha})$ will have expectation close to zero. The variances of these statistics are generally difficult to approximate and involve fourth moments of terms in $\log y_{ij}$.

Instead of approximating these variances, we will look at individual terms, *i.e.*

$$D_{ij} = \psi' \left(\sum_i \widehat{\alpha}_i \right) - \psi' (\widehat{\alpha}_i) + s_{ij}^2$$

as diagnostics of overdispersion in α_i of individual observations.

When the Dirichlet model includes covariates instead of α_i we have

$$\alpha_{ij} = \exp \{ \beta'_i z_j \}$$

and the overdispersion statistics become

$$D_{ij} = \psi' \left(\sum_i \widehat{\alpha}_{ij} \right) - \psi' (\widehat{\alpha}_{ij}) + s_{ij}^2$$

and

$$D_{ij}(\beta_{il}) = a_{ij}^2 - z_{jl}^2 D_{ij}$$

for tests of homogeneity of the parameters a_{ij} and the individual regression coefficients β_{il} respectively. **All of these D_{ij} statistics** are the difference of two measures of the observed information for the estimated variance of a parameter estimate, as described in the Appendix.

5.2 Global and Local Influence

The Cook's Distance Δ_j associated with the j -th observation is the change in the maximized log-likelihood function when the j -th observation is deleted (jackknifed) and then the model is refitted. Intuitively, Δ_j represents the global influence that the j -th observation has on the overall fit of the model. Cook (1977) is concerned with linear regression with independent normal errors but his ideas continue to hold in more general settings such as the present.

In Fig. 6 we see that Cook's distance Δ_j is highly correlated with the composite residual C_j in these data. These composite residuals are also highly correlated with the Mahalanobis M_j as well. Intuitively, the Dirichlet log-likelihood function is approximately the same as that of a multivariate normal, which, in turn, is approximately the sum of quadratic functions of the raw residuals d_j after making allowances for the differing variances between individuals. Our interest is not in remarking on these correlations between Δ_j , C_j , and M_j , but rather, in identifying the extremes of these metrics that allow us to identify unusual observations.

Cook (1986) suggests that local measures of influence are obtained by giving an observation slightly less weight in the likelihood rather than deleting it completely, as with Cook's distance. The details specific to the present application are given in the Appendix. These measures of local influence are standardized score statistics defined as

$$\delta_{ij} = s_{ij} / \left\{ \psi'(\hat{\alpha}_{ij}) - \psi' \left(\sum_i \hat{\alpha}_{ij} \right) \right\}.$$

This is the local influence that the j -th subject has on the parameter estimates. The denominator of δ_{ij} is the amount of observed information that y_{ij} contributes to the estimation of $\hat{\alpha}_{ij}$. Notice the similarities in functional forms of D_{ij} and δ_{ij} .

6 Specific application

6.1 Analysis of total PANSS

The primary analysis seeks to explain total PANSS in terms of baseline variables. We performed a linear regression using least squares and all available covariates described in Section 1. The only explanatory variable that achieves a high level of statistical significance is the Simpson-Angus ($\hat{\beta}=4.48$, $SE = 1.01$). To a much smaller degree, the duration of illness exhibits a modest amount of explanatory value ($\hat{\beta}=0.274$, $SE = 0.140$), with longer durations associated with larger (*i.e.* worse) values of total PANSS. This seemingly important covariate value was missing in 13% of the subjects but missingness appears to be non-informative: A simple frequency analysis showed that whether or not an individual was missing this covariate value was unrelated to whether he scored above or below the sample mean on PANSS ($\chi^2 = 0.43$; 1df; $p = .51$). Even though the duration of illness values were missing in some subjects, we chose to keep this important covariate and perform the

subsequent statistical analysis of the components of PANSS on the sample of 358 subjects with complete data.

The overall F ratio for the regression is 5.18 and (7, 350) df with $p < .0001$ indicating good explanatory value of the combined baseline covariates on describing the overall behavior of total PANSS. We next examine the five individual components of the PANSS.

6.2 Component regression

Fig. 1 plots the marginal frequencies of each of the component scores $x_1, \dots, x_5$. We see that $x_1, \dots, x_4$ all follow roughly symmetric, unimodal distributions, with modes bounded away from the lowest limits of their range. The distribution of x_5 is skewed with its mode at the lower limit of its range. A comparison of the marginal means and standard deviations in Table 1 shows that the variances of all fractional components $y_1, \dots, y_5$ are larger for those components that make a larger contribution to the total PANSS, as anticipated by the Dirichlet distribution. This is the first indication that the Dirichlet distribution is appropriate for these data. Another indication is that all but one of the correlations among the components are negative as predicted by the Dirichlet distribution. The only non-negative correlation is $r=0.01$ and it is not statistically significant ($p=0.86$)

Estimates of the fitted β 's in (2.4) are examined next. We fitted linear terms in all explanatory baseline covariates (except sex because the few female subjects were omitted) plus total PANSS and intercepts, for each of the five components, for a total of 45 regression parameters. Compared to a model with only five intercepts, the likelihood ratio chi-squared statistic for this model had a significance level of $p < 10^{-6}$ indicating a high degree of overall explanatory value. Since each variable was statistically significantly associated with at least one component, we kept this model as the final model. Parameter estimates are summarized in Table 2.

In order to facilitate comparisons of regression parameters between components when there were more than two significant regression coefficients, we coded an overall regression slope β_{0l} on l -th covariate and then individual slopes $\beta_{2l}, \dots, \beta_{5l}$ in each of expression (2.4) of $\alpha_2, \dots, \alpha_5$. The regression slope in α_1 was then expressed as $\beta_{0l} - \beta_{2l} - \dots - \beta_{5l}$. This type of coding was used for the effect of the total PANSS since total PANSS was significantly associated with each of the individual components. Hence the parameter for the effect of total PANSS on positive symptoms in Table 2 is the negative of the sum of the parameters of the remaining four components.

Overall, total PANSS proved highly statistically significant in explaining each of the five α_i values, both overall as well as differentially for each of the five individual components. Among the other covariates available to us, the Simpson-Angus contributed significantly to two components, while history of depression, history of substance abuse, and AIMS each contributed to one component. Increasing total PANSS was associated with a diminishing proportion of positive and emotional symptoms, and with increasing proportion of negative, cognitive and hostility symptoms. Parkinsonian side effects (higher Simpson-Angus scores) were associated with relative worsening of negative and emotional symptoms while higher Tardive Dyskinesia scores (on the AIMS scale) were associated with relative worsening of the cognitive component. Lifetime history of depression was associated with relative worsening of emotional symptoms. Lifetime substance abuse was associated with higher meta-parameters α_i for all components but there were no significant differences among components. Hence the relative contributions of all components did not significantly change with this covariate.

Fig. 2 plots the fitted expected values for each of the five component values along with the fitted estimate of $\sum \alpha_i$ versus total PANSS (the only significant covariate for all five components). All other covariates assume their average values in this figure. The increasing value of $\sum \hat{\alpha}_i$ in this figure also indicates a decreasing variance in each of the individual components for larger values of total PANSS.

At the most extreme possible values of total PANSS of 29 and 203, all Y_i have means determined by their ranges and zero variances so modeling will fail in extrapolating the Dirichlet regression model to include these extreme limits. These limits are far beyond the range of Fig. 2 as well as the range of the observed data.

6.3 Model diagnostics

Fig. 6 is a scatter plot of the composite residuals C and Cook's distances Δ for all subjects under the fitted model given in Table 2. There is a strong correlation between these two diagnostic measures as noted in Section 2. More importantly the three unusual subjects that are indicated with *'s in this figure are most extreme according to both the composite residual and Cook's distance. A similar pattern appeared when we plotted either of these measures against the Mahalanobis M (including the same three outliers) so these figures are not included. The three outliers in this figure appear more striking on the C axis than the Δ axis indicating that they are more likely to be classified by the viewer as "outliers" when compared to the other values of C . The plot of M values is closer to C than Δ in identifying these three as outliers. The star and tree plots of the three observations with the largest composite residuals (not shown) did not reveal any systematic pattern.

Table 3 shows the values of the overdispersion statistics D_{ij} . These are expressed as values of $t = \text{mean}/\text{SEM}$ with averages taken over index j on subjects. Positive values are indicative of overdispersion as measured by the difference in two estimates of the observed information. Large negative values indicate either a very good fit or indicate problems elsewhere.

Overdispersion in α_2 (PANSS Negative) is suspected, with a t -value of 2.29 in Table 3 but none of the individual regression coefficients, including the intercept, indicate the source of the problem. With the exception of overdispersion of the regression coefficient in Depression (which is not included in the final model in Table 2), all of the other covariates appear to contribute roughly equally to this lack of fit. PANSS Hostility (y_5) exhibits overdispersion in several regression coefficients, but not in terms of the overall α_5 metaparameter. Overdispersion can appear in regression coefficients not included in the model and may indicate that the effect of the covariate is not linear in (2.4). We next look at Y_2 more closely.

Fig. 7 plots individual r_2 PANSS Negative residuals against the measure of overdispersion of the meta-parameter α_2 . The plot character is the value of x_2 rounded to the nearest 10. Large negative residuals in r_2 are associated with low values in x_2 of 1 or 2, and large positive residuals are associated with medium to high values of x_2 indicating that the Dirichlet model has difficulty fitting the more extreme observed values. Larger values of the overdispersion measure D_{ij} correspond to these extremes, which are also influential in the fitting. The three outliers identified in Fig. 6 are depicted with *'s here also. While two of these three have large absolute r_2 residuals, they are not among the four extreme α_2 overdispersion observations separated in the upper left corner of this plot.

A plot of D_{ij} against either r_{ij} , s_{ij} , or δ_{ij} will generally take a “U” shape as illustrated in Fig.

7. Intuitively, D_{ij} differs by a constant from s_{ij}^2 and we found that r_{ij} , s_{ij} , and δ_{ij} are all very highly correlated with each other.

In conclusion, the Dirichlet model offers a good overall fit to the data. The three extreme composite outliers were identified by each of the C , M and Cook’s distance Δ composite residuals. Overdispersion in PANSS Negative (Y_2) can be traced to the Dirichlet model’s inability to fit extreme values of the original scores X_2 that define this component.

From the clinical point of view some of the findings of the Dirichlet regression modeling are novel and informative. Most notably, we found that increases in PANSS total scores are associated with increased negative symptoms, cognitive dysfunction and hostility and with decreased positive and general symptom scores. This result suggests that the illness presents with a basic floor of positive symptoms and that as the illness gets more severe, its negative, cognitive and hostility symptoms add to increased severity.

Acknowledgments

We wish to thank Dr. John Krystal for helpful discussions.

Funding: Grant MH62294 awarded to the Yale Center for Interdisciplinary Research in AIDS.

Appendix: Local Jackknife and Overdispersion Diagnostics

We briefly describe the diagnostic measures used here first in general terms and then for the Dirichlet model in particular. For an independent sample $\{x_j, j = 1, \dots, n\}$ of random variables $\{X_j\}$ from the same parent population with density function $f(x | \theta)$, denote the log-likelihood function $\Lambda(\theta)$ that depends on a real-valued parameter θ and write

$$\Lambda(\theta) = \sum_j^n \log f(x_j | \theta).$$

The score function

$$S(\theta) = \partial \Lambda(\theta) / \partial \theta = \sum s(x_j | \theta)$$

satisfies $S(\hat{\theta}) = 0$ where $\hat{\theta}$ is the maximum likelihood estimate and

$$s(x_j | \theta) = \partial \log f(x_j | \theta) / \partial \theta.$$

The observed information for θ is defined as

$$I(\theta) = \sum_j i(x_j | \theta) = - \partial^2 \Lambda(\theta) / \partial \theta^2$$

where

$$i(x_j | \theta) = - \partial^2 \log f(x_j | \theta) / \partial \theta^2.$$

The statistics s , I , and i are always evaluated at $\theta = \widehat{\theta}$.

The measure of overdispersion of θ described by Zelterman and Chen (1988)

$$D(\widehat{\theta}) = \sum_j s^2(x_j|\widehat{\theta}) - i(x_j|\widehat{\theta})$$

is the difference of two estimates of the observed information.

The local jackknife estimate $\widehat{\theta}_{(j)}$ due to Cook (1986) is the maximum likelihood estimate of θ obtained by giving observation x_j a weight of $1 - \epsilon$ for a value of ϵ very close to zero and all other observations are given a weight of 1.

That is, $\widehat{\theta}_{(j)} = \widehat{\theta}_{(j)}(\epsilon)$ is the solution in θ of the equation

$$S(\theta) - \epsilon s(x_j|\theta) = \theta$$

For values of ϵ very close to zero, $\widehat{\theta}_{(j)}$ will be close to $\widehat{\theta}$. The solution to this last equation can then be approximated using a one-step Newton iteration giving

$$\widehat{\theta}_{(j)} \approx \widehat{\theta} - \epsilon s(x_j|\widehat{\theta}) / \{I(\widehat{\theta}) - \epsilon i(x_j|\widehat{\theta})\}.$$

In most settings, $I(\widehat{\theta})$ is $O(n)$ so that the i in this denominator can be ignored.

For values of ϵ close to zero, the diagnostic

$$\delta_j(\widehat{\theta}) = s(x_j|\widehat{\theta}) / I(\widehat{\theta})$$

can be interpreted as the amount of local influence that the observation x_j contributes to the estimation of $\widehat{\theta}$.

In the specific application to Dirichlet models without covariates, begin at (2.1) and write

$$\Lambda(\alpha) = \sum_j^n \left\{ \log \Gamma \left(\sum_i \alpha_i \right) - \sum_i \log \Gamma(\alpha_i) + \sum_i (\alpha_i - 1) \log y_{ij} \right\}$$

where y_{ij} is the i -th component ($i = 1, \dots, k$) measured on the j -th subject ($j = 1, \dots, n$).

The score residual for α_i is

$$s(\alpha_i) = \psi \left(\sum \widehat{\alpha}_i \right) - \psi(\widehat{\alpha}_i) + \log y_{ij}$$

where $\psi(\theta) = \partial \log \Gamma(\theta) / \partial \theta$. We will use this notation for s rather than the more accurate but awkward $s_{\alpha_i}(\widehat{\alpha})$.

Similarly,

$$i(\alpha_i) = \psi'(\hat{\alpha}_i) - \psi'(\sum \hat{\alpha}_i)$$

where $\psi'(\theta) = \partial^2 \log \Gamma(\theta) / \partial \theta^2$. Again, we employ a simplified notation.

The measure of overdispersion in α_i due to the j -th observation is then

$$D_{ij}(\alpha_i) = \psi'(\sum \hat{\alpha}_i) \psi'(\hat{\alpha}_i) + \left\{ \psi(\sum \hat{\alpha}_i) - \psi(\hat{\alpha}_i) + \log y_{ij} \right\}^2$$

and the local influence of the j -th observation on α_i is

$$\delta_{ij}(\alpha_i) = \left\{ \psi(\sum \hat{\alpha}_i) - \psi(\hat{\alpha}_i) + \log y_{ij} \right\} / n \left\{ \psi'(\sum \hat{\alpha}_i) - n \psi'(\hat{\alpha}_i) \right\}.$$

To examine diagnostics concerning the individual regression coefficients β instead of α_i write

$$\alpha_{ij} = \exp \{ \beta'_i z_j \}$$

from (2.4) where $z_j = \{z_{jl}\}$ and $\beta_i = \{\beta_{il}\}$ for covariates $l = 1, \dots, L_i$ measured on the i -th component of the j -th subject.

Then write

$$\partial \alpha_{ij}(\beta_i) / \partial \beta_{il} = z_{jl} \alpha_i(\beta_i).$$

Use the chain rule to obtain score residuals

$$s_{ij} = \alpha_{ij} - z_{jl} s(\alpha_{ij})$$

and overdispersion statistics

$$D_{ij}(\beta_{il}) = \alpha_{ij}^2 z_{jl}^2 D_{ij}(\alpha_{ij})$$

for each of the regression coefficients in the model (2.4).

References

- Abramowitz, M.; Steegun, IA. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. J. Wiley & Sons; New York: 1972.
- Aitchison, J. The Statistical Analysis of Compositional Data. Chapman & Hall; New York: 1986.
- Aitchison J, Shen SM. Logistic-normal distributions: Some properties and uses. Biometrika. 1980; 67:261–72.
- Barcelo C, Pawlowsky V, Grunsky E. Some aspects of transformation of compositional data and the identification of outliers. Mathematical Geology. 1996; 28:501–18.

- Baxter MJ. Detecting multivariate outliers in artefact compositional data. *Archaeometry*. 1999; 41:321–38.
- Bell MD, Lysaker PH, Beam-Goulet JL, Milstein RM, Lindenmayer JP. Five-component model of schizophrenia: assessing the factorial invariance of the positive and negative syndrome scale. *Psychiatry Research*. 1994; 52:295–303. [PubMed: 7991723]
- Campbell G, Mosimann JE. Multivariate methods for proportional shape. *ASA Proceedings of the Section on Statistical Graphics*. 1987:10–17.
- Connor RJ, Mosimann JE. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*. 1969; 64:194–206.
- Cook RD. Detection of influential observations in linear regression. *Technometrics*. 1977; 19:15–8.
- Cook RD. Assessment of local influence (with discussion). *Journal of the Royal Statistical Society*. 1986; 48:133–69.
- Espinheira PL, Ferrari SLP, Cribari-Neto F. On beta regression residuals. *Journal of Applied Statistics*. 2008; 35:407–19.
- Espinheira, PL.; Ferrari, SLP.; Cribari-Neto, F. Computational Statistics and Data Analysis. 2008. Influence diagnostics in beta regression. Available online at <http://www.sciencedirect.com/science/journal/01679473>
- Ferrari S, Cribari-Neto F. Beta regression for modeling rates and proportions. *Journal of Applied Statistics*. 2004; 31:799–815.
- Hijazi, RH.; The American University. Ph.D. Dissertation. Washington, D.C.: 2003. Analysis of compositional data using Dirichlet covariate models.
- Hijazi, RH. Technical report: Department of Statistics. United Arab Emirates University; 2006. Residuals and diagnostics in Dirichlet regression.
- Janno S, Holi MM, Tuisku K, Wahlbeck K. Validity of Simpson-Angus Scale (SAS) in a naturalistic schizophrenia population. *BMC Neurology*. 2005; 5:5. [PubMed: 15774006]
- Kay SR, Opler LA, Lindenmayer JP. The Positive and Negative Syndrome Scale (PANSS): rationale and standardisation. *British Journal of Psychiatry*. 1989; 7:59–67.
- Rao, CR. Linear Statistical Inference and its Applications. Wiley; New York: 1973.
- Rosenheck R, Cramer J, Xu W, Thomas J, Henderson W, Frisman L, Fye C, Charney D. A comparison of clozapine and haloperidol in hospitalized patients with refractory schizophrenia. *New England Journal of Medicine*. 1997; 337:809–15. [PubMed: 9295240]
- Shen SM. A method for discriminating between models describing. *Biometrika*. 1982; 69:587–95.
- Simpson GM, Angus JWS. A rating scale for extrapyramidal side effects. *Acta Psychiatrica Scandinavica*. 1970; 212:11–9.
- Zelterman D, Chen C-F. Homogeneity tests against central mixture alternatives. *Journal of the American Statistical Association*. 1988; 83:179–82.

Figure 1.
Observed marginal frequencies $x_1, \dots, x_5$ of the individual components of the PANSS score.

Figure 2. Fitted expected values varying total PANSS using fitted regression parameters given in Table 2 using the left scale. Other covariates are fixed at their average values. The estimated value $\sum \alpha_i$ is plotted in a dashed line using the scale on the right.

Figure 3.

Example of a mosaic plot of PANSS fractional components for the seven individuals with the largest Simpson-Angus, sorted by values of y_1 . Values of their PANSS scores are given at the bottom.

Figure 4.
Tree plot of residuals for the individual with the largest composite residual C_j .

Figure 5.
Star-plot of residuals for the same individual depicted in Fig. 4

Figure 6.
Composite residuals C and Cook's distance Δ .

Figure 7. Overdispersion in α_2 plotted against the PANSS negative subscale component r_2 standardized residual. The values plotted are the x_2 PANSS negative component rounded to the nearest 10. The three outliers in are *'ed here.

Labels, interpretations, valid ranges, marginal means, and standard deviations of the five component values X_i of the PANSS total score and its fractional components Y_i .

Table 1

Variable index i	Subscale interpretation	Valid range	Component x_i		Fraction of total y_i	
			Mean	SD	Mean	SD
1	Positive	6-42	24.09	4.18	0.272	0.045
2	Negative	8-56	23.20	6.30	0.258	0.045
3	Cognitive	7-49	21.78	5.36	0.243	0.042
4	Emotional	4-28	12.13	3.25	0.137	0.037
5	Hostility	4-28	8.07	3.15	0.090	0.031
PANSS score			89.27	13.35		

Table 2

Estimated regression parameters for final 12 parameter model, followed by standard errors, and t-values. The fitted means of $Y_1, \dots, Y_5$ from this model are plotted in Fig. 2.

Covariate	All	Component				
		Positive Y_1	Negative Y_2	Cognitive Y_3	Emotional Y_4	Hostility Y_5
Intercept	1.347	.729	-.156	.119	.107	-.800
	.374	.094	.095	.097	.120	.146
	3.60	7.74	-1.65	1.23	.90	-5.49
Total	.016	-.004	.004	.003	-.005	.003
PANSS	.003	.0007	.0007	.0007	.0009	.001
	5.23	-5.51	5.11	3.44	-5.60	2.69
Depression	.448	-.014	.063	-.065	.130	-.114
	.148	.032	.032	.034	.039	.050
	3.02	-.46	1.99	-1.94	3.37	-2.29
Substance	.206	.006	-.048	-.046	.013	.074
abuse	.085	.023	.023	.023	.029	.034
	2.41	.28	-2.07	-2.00	.46	2.17
Simpson-	.043	-.018	.033	-.026	.058	-.046
Angus	.058	.014	.015	.015	.019	.022
	.73	-1.27	2.16	-1.72	3.07	-2.12
Length of	-.007	.003	-.005	.005	-.004	.001
illness	.008	.002	.002	.002	.002	.003
	-.84	1.82	-2.56	2.22	-1.63	.35
Age	.003	-.0004	.005	-.003	.001	-.003
	.009	.002	.002	.002	.003	.003
	.38	-0.21	2.60	-1.16	.40	-1.12
Race	.019	-.019	.046	-.020	-.019	-.027
	.085	.021	.022	.022	.027	.032
	.22	-0.92	2.08	-.89	-.68	-.84
AIMS	-.016	-.002	-.0007	.006	-.0002	-.004

Covariate	Component				
	Positive		Negative		Hostility
	All	Y ₁	Y ₂	Cognitive Y ₃	Emotional Y ₄
	.008	.002	.002	.002	.002
	-2.05	-.88	-.35	3.29	-.09
					.003
					-1.32

Table 3 Measures of overdispersion statistics expressed as $t = \text{mean/SEM}$.

Overdispersion statistic	PANSS Component			
	Positive	Negative	Cognitive	Emotional Hostility
α	-1.85	2.08	-2.06	0.44 0.92
Dirichlet regression parameters:				
Intercept	-2.24	1.37	-2.73	0.33 2.03
Total PANSS	-2.73	0.93	-2.99	0.29 2.31
Depression	-1.68	-0.64	-0.84	0.98 -1.24
Substance abuse	-1.05	1.06	-1.60	0.14 1.49
Simpson-Angus	-1.20	0.81	-2.76	0.65 1.84
Length of illness	-2.85	1.31	-4.91	0.84 2.07
Age	-2.40	1.49	-3.73	0.67 2.04
Race	-1.33	1.12	-0.88	0.07 -0.13
AIMS	-2.00	0.61	-2.32	1.16 0.78