Spherical Subfamily Models

Alan Gous
Stanford University

November 10, 1999

Abstract

A new method is presented for modeling low-dimensional representations of high-dimensional multinomial and compositional data. The data are fit to subfamilies of the multinomial family which are defined using the multinomial information geometry. These collections of *spherical subfamilies* have a number of advantages over the affine subfamilies contructed by methods such as canonical and correspondence analysis, traditionally fit to such data. First, they can describe more complex shapes in the data, and are particularly well-suited to modelling sparse data. Second, the subfamilies provide a convenient variance-stabilizing parametrization for the fitted data. An algorithm which uses iterative Singular Value Decompositions is presented for fitting the models.

Two example applications are presented: one is to Latent Semantic Indexing, a method for the automatic indexing of text documents. The ability of the method to model sparse data is an advantage here. A second example is an analysis of compositional data from a geological study. This example shows the ability of the method to model curvature in the data, and illustrates its variance stabilization properties.

Keywords: Multinomial; Compositional data; Information geometry; Latent Semantic Indexing; Singular Value Decomposition; Sparse data.

1 Introduction

The Singular Value Decomposition (SVD) is a standard tool used to provide a low-dimensional representation of high-dimensional data. Truncating the SVD of the data matrix projects the data onto the best-fitting affine subspace of a specified dimension. Such a representation may be used to help visualize and interpret the data (for example in Principal Component Analysis), or simply to reduce variance in estimation. The method has statistical justification if we are prepared to assume that the data are multivariate normal, with mean vectors lying on just such an affine subspace. In this case the procedure is maximum likelihood.

Multinomial data, such as those in a contingency table, may be modeled in a similar way. Canonical analysis (Hirschfield, 1935) uses reduced rank approximations to such a table, effectively modeling

the probability vectors generating such data as lying on a low-dimensional affine subspace of the multinomial simplex. Correspondence analysis (Greenacre, 1984) fits these reduced-rank models using a truncated SVD of the contingency table, appropriately transformed to take some account of the multinomial variance structure. Greenacre and Hastie (1987) emphasize the geometric interpretation of this procedure. Canonical Analysis receives a formal maximum likelihood treatment in Gilula and Haberman (1986).

But certain multinomial data do not appear to be modelled well by affine restrictions on the underlying probability vectors. One example is large, sparse matrices of counts. These occur, for example, in modeling the frequencies of word occurrence in collections of text documents.

We introduce a new class of low-dimensional manifolds, taylored specifically for multinomial data. These *spherical subfamilies* are shaped very differently to the affine subspaces in the multinomial simplex. In Section 2 we analyse a large, sparse dataset of word counts in a document. We show that spherical subfamilies provide better models for the data and, in particular, model the sparsity of the data substantially better than the affine models.

These spherical subfamilies have another advantage too. They are defined using the *information geometry* of the multinomial family, and there is a convenient parametrization of the subfamilies which is variance stabilizing. Using this parametrization, distances between data points fit to these subfamilies may be compared on an approximately constant scale. In Section 3 we illustrate this property by analysing a simple geological dataset. This is compositional, not multinomial, data, but the variance structure of multinomial data is a reasonable model for the variance structure of these data, and with this model the variance-stabilizing property is retained.

Spherical subfamily models are defined and fit using similar methods as are used to define and fit affine subspace models. Section 4 reviews affine subspace fits for multinomial models, formally defining the *Affine Subfamily Models*. These are very similar to canonical models, but are defined from a more geometric point of view. Section 5 describes the variance-stabilizing properties of the multinomial information geometry, and Sections 6 and 7 define the *Spherical Subfamily Models*, formalizing the methods used in Sections 2 and 3. Lastly, Section 8 describes an iterative algorithm for fitting these models.

2 Example: Latent Semantic Indexing

Latent Semantic Indexing (LSI, see Berry et al., 1995) is an automated method of indexing collections of documents stored in electronic form. The method represents documents as points in a low-dimensional space calculated using the frequencies of word occurrence in each document. The relative distances between these points are interpreted as distances between the topics of the documents, and can be used to find related documents, or documents matching some specified query.

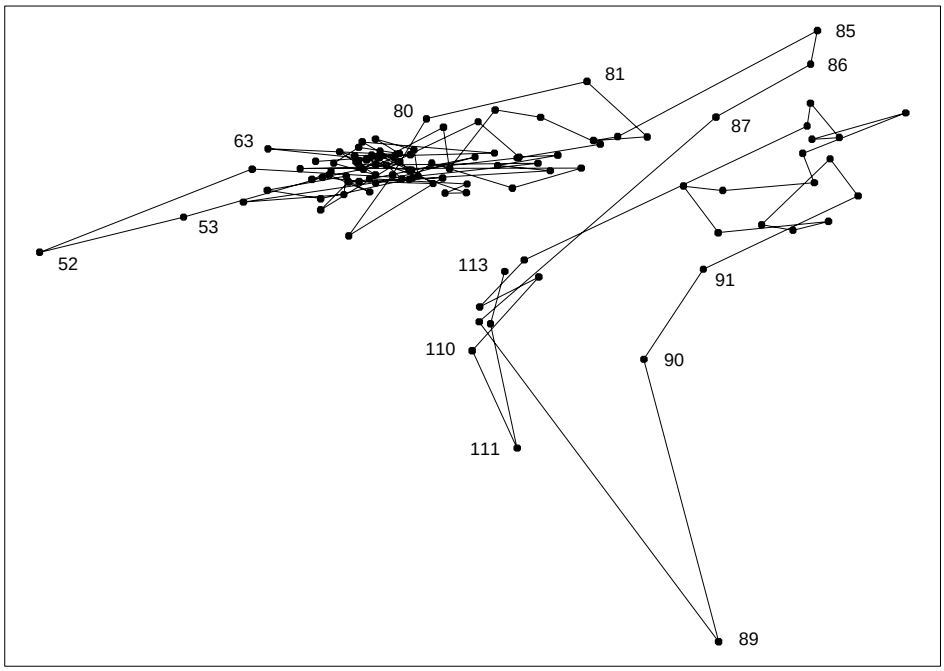


Figure 1: The 113 pages of *The Prince* projected onto a two-dimensional space using an Affine Subfamily Model, fit by minimizing the approximate Chi-squared distance (8). Consecutive pages are connected by lines; some of the pages numbers are shown.

The documents in the collection could be parts of a larger document, such as pages in a book. As an example, we will consider the English translation of Machievelli's *The Prince*, available on the Internet at `http://www.gutenberg.net`. The text consists of 45155 words, of which 4659 are unique. We will define 113 "pages" of this text to be 112 blocks of 400 consecutive words each, and one last block of 355 words. Figure 1 shows an extreme example of dimension reduction. An Affine Subfamily Model (ASM, described below and defined formally in Section 4) has been used to represent each of the 113 pages as a point on a two-dimensional plane. Successive pages have been connected by lines, so tracking the evolution of the subject matter of the book over the course of its pages. The evolution appears smooth in places, and there is a suggestion of a change in topic around page 85. In fact, the first 85 or so pages deal with political theory and the last part is more biographical. So the break suggested by the diagram does signify a real change in the subject matter.

In this section we will show how LSI may be framed as a statistical estimation problem. We will then compare the performance of two different classes of dimension-reduction models, using the *Prince* data as a test case. The first are the ASMs mentioned above. These are the models traditionally used in LSI. The second are the Spherical Subfamily Models, which we will describe briefly here and in greater detail in Sections 6 and 7. We will show that the *Prince* data is modeled far better using this second class of models.

To proceed more formally, suppose we have K documents numbered $1, \ldots, K$, and $p+1$ index terms numbered $1, \ldots, p+1$. In the above example $K = 113$ and $p+1 = 4659$, since we will use each unique word as an index term. For each $i = 1, \ldots, K$ let x_i be a $(p+1)$-vector with jth element x_i^j equal to the number of times term j appears in document i, $j = 1, \ldots, p+1$. (Note: since we will be using subscripts to enumerate vectors themselves, we will use superscripts throughout this paper to refer to elements within vectors.)

Our model for the "topic" of each document i will be a $(p+1)$-vector of probabilities π_i, with jth element π_i^j the probability that a term picked at random from document i, or from a document on the same topic, is term j. We will model the x_i as independent random variables with

$$x_i \sim \text{Multinomial}_p(n_i, \pi_i), \quad n_i = \textstyle\sum_j x_i^j, \quad i = 1, \ldots, K. \tag{1}$$

The note at the end of the section discusses this choice of model.

Without any further assumptions on the parameters, the maximum likelihood estimates (MLEs) of the π_i are

$$\hat{\pi}_i = x_i / n_i, \quad i = 1, \ldots, K. \tag{2}$$

The π_i and $\hat{\pi}_i$ lie in the p-dimensional multinomial simplex

$$E_p = \{\pi = (\pi^1, \ldots, \pi^{p+1}) \in \mathbb{R}^{p+1},\ \textstyle\sum_{j=1}^{p+1} \pi^j = 1,\ \pi^j \geq 0 \text{ for each } j\}. \tag{3}$$

Referring back to the *Prince* data above, since most of the 113 pages in the book will contain very few of the 4659 words, and will seldom contain any one of the words more than once or twice, the x_i in (1) will be very sparse, and the $\hat{\pi}_i$ in (2) will have most entries very close, or equal, to zero. So the $\hat{\pi}_i$ will tend to lie on or near the faces of the simplex E_p rather than the center.

Because of this, as well as for semantic reasons, it seems reasonable to assume that not all elements of E_p will be appropriate descriptions of a "topic" of a document. We expect that the $\pi \in E_p$ describing realizable documents lie in some much smaller subset of E_p.

We certainly do not know *a priori* what subset of E_p would be appropriate for modeling topics of documents. Instead we can specify a collection of subsets, and include the choice of subset as part of the estimation problem. Each point $\pi \in E_p$ corresponds to a distribution in the family of p-dimensional multinomial distributions

$$\mathcal{M}_p = \{\text{Multinomial}_p(1, \pi) : \pi \in E_p\}. \tag{4}$$

A subset of E_p corresponds to a *subfamily* $\mathcal{F}$ in $\mathcal{M}$. We will use expressions such as "the point $\pi \in \mathcal{F}$" as shorthand for "the family $\text{Multinomial}_p(1, \pi) \in \mathcal{F}$".

So we specify a collection F of subfamilies $\mathcal{F}$ of $\mathcal{M}_p$, and assume all π_i lie in one $\mathcal{F} \in F$. It is easy to see that with this restriction on the π_i their MLEs, which we will call $\tilde{\pi}_i$, minimize

$$\sum_{i=1}^{K} n_i \text{Dev}(\hat{\pi}_i, \tilde{\pi}_i) \tag{5}$$

over $\tilde{\pi}_i \in \mathcal{F}$ and $\mathcal{F} \in F$. Here Dev is the *deviance* function in the multinomial family,

$$\mathrm{Dev}(\pi_1, \pi_2) = 2 \sum_{j=1}^{p+1} \pi_1^j \log \left(\frac{\pi_1^j}{\pi_2^j} \right), \quad \pi_1, \pi_2 \in E_p. \tag{6}$$

The deviance (see McCullagh and Nelder, 1989) is twice the Kullback-Leibler divergence.

It remains to specify F, and then to give some algorithm for minimizing (5), or some approximation to it. A useful form of F is some collection of q-dimensional *submanifolds* of E_p, for some fixed q, $0 \leq q < p$. In particular, we can choose F to be the collection of non-empty intersections of E_p with q-dimensional affine subspaces of $\mathbb{R}^{p+1}$. Call this collection $F_A^{p,q}$, or just F_A when the two dimensions are understood. This defines the (p, q)-Affine Subfamily Model, or (p, q)-ASM.

The *Prince* dataset is very large, and typical LSI applications are even larger. For ease of computation then we will fit these models not using the iterative maximum likelihood fitting algorithm such as that outlined in Gilula and Haberman (1986), but instead use an approximate fitting criterion which is more easily optimized. The chi-squared distance approximation to (5) is

$$\sum_{i=1}^{K} \sum_{j=1}^{p+1} n_i \frac{(\hat{\pi}_i^j - \tilde{\pi}_i^j)^2}{\tilde{\pi}_i^j}, \tag{7}$$

which we will approximate further by

$$\sum_{i=1}^{K} \sum_{j=1}^{p+1} n_i \frac{(\hat{\pi}_i^j - \tilde{\pi}_i^j)^2}{\hat{\pi}_0^j}, \tag{8}$$

where $\hat{\pi}_0$ is (in the Correspondence Analysis terminology) the *centroid* of the $\hat{\pi}_i$,

$$\hat{\pi}_0 = \frac{\sum_i x_i}{\sum_i n_i}. \tag{9}$$

Section 4 describes how this approximation may be minimized using an SVD, in a manner very similar to Correspondence Analysis.

Choosing $q = 2$, the ASM estimates $\tilde{\pi}_i$ lie on a two-dimensional plane, and it is this plane which has been represented in Figure 1. Practitioners of LSI typically use dimensions q in the hundreds when p, as in this example, is in the thousands.

We will briefly describe another collection of submanifolds of E_p here, deferring a more formal definition to Section 6. This collection is defined using the *spherical parameterization* of the p-dimensional multinomial family, obtained by the transformation

$$\theta = \theta(\pi) = (\sqrt{\pi^1}, \ldots, \sqrt{\pi^p}), \quad \pi \in E_p. \tag{10}$$

The parameters θ lie on the positive orthant of the p-dimensional unit sphere S_p,

$$S_p^+ = \{\theta : \sum_j (\theta^j)^2 = 1, \ \theta^j \geq 0 \text{ for each } j\}. \tag{11}$$

The natural distance measure between points θ_1 and θ_2 on the unit sphere is the geodesic distance, $\arccos(\theta_1^T\theta_2)$. A useful feature of the spherical parameterization is the relationship between multinomial deviance and the geodesic distance on S_p^+:

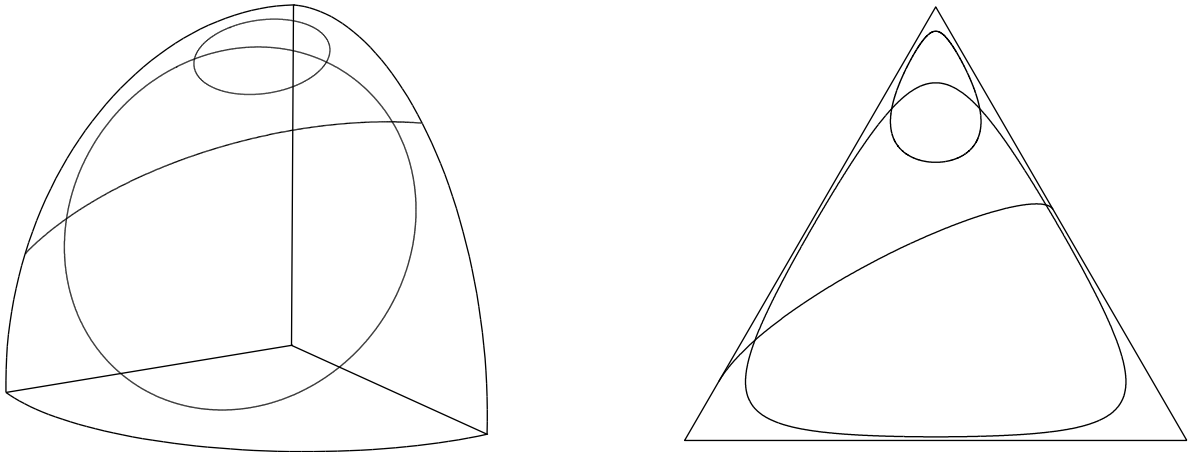
$$\mathrm{Dev}(\theta_1, \theta_2) = d(\theta_1, \theta_2)^2 + o(d(\theta_1, \theta_2)), \tag{12}$$

where d, called the *information distance* in the family $\mathcal{M}_p$, is twice the geodesic distance:

$$d(\theta_1, \theta_2) = 2\arccos(\theta_1^T\theta_2), \quad \theta_1, \theta_2 \in S_p^+. \tag{13}$$

Section 5 gives more details about this distance function.

For any q, $p \geq q \geq 1$, there are q-dimensional *subspheres* of S_p. These are q-dimensional spheres lying within S_p, and may have any radius between 0 and 1. We define the (p, q)-*spherical subfamilies* to be the q-dimensional subfamilies of $\mathcal{M}_p$ corresponding to the collection of q-dimensional subspheres of S^p, intersected with S_+^p. We will denote this collection of subfamilies $F_S^{p,q}$, or just F_S.



(a) (b)

Figure 2: Examples of spherical subfamilies. The positive orthant of the sphere S_2 in (a) is mapped onto E_2 in (b) using the spherical parameterization. Three subspheres are shown in (a), with radii 0.2, 0.5 and 1. The corresponding one-dimensional submanifolds of E_2 are shown in (b).

Figure 2 shows the trinomial simplex E_2 (a triangle) and the corresponding orthant S_2^+. Three 1-dimensional subspheres (circles) of S_2^+ are shown. These are members of $F_S^{2,1}$. One of these has radius 1, and only an arc of the circle lies in S_2^+. The radii of the others are both less than 1, and they lie completely in S_2^+. The corresponding submanifolds in E_2 are shown.

The arc with radius 1 is almost a straight line in E_p, and so is a very similar shape to a particular subfamily in $F_A^{2,1}$. The other two subspheres however are of a completely different shape. Note

that both of them trace out curves very close to the boundary of the simplex. It is subfamilies such as these, in much higher dimensions, which can model well the sparsity of document data.

The (p, q)-Spherical Subfamily Model (SSM) restricts data as in (1)–(2) to lie in an unspecified $\mathcal{F} \in F_S^{p,q}$. Put $\hat{\theta}_i = \theta(\hat{\pi}_i)$ for each i. Motivated by (12) we estimate this $\mathcal{F}$ by minimizing an approximation to (5), the sum of squared information distances

$$\sum_{i=1}^{K} n_i d(\hat{\theta}_i, \tilde{\theta}_i)^2, \tag{14}$$

over $\tilde{\theta}_i$ all restricted to lie on one $\mathcal{F} \in F_S$. Note that minimizing (14) is equivalent to minimizing the corresponding sum of squared geodesic distances in S_+^p.

Section 8 describes an iterative algorithm for fitting SSMs by minimizing (14). Each step in the iteration involves an SVD calculation. The first step of the iteration provides an approximation to this fit which, requiring the calculation of just one SVD, is computationally comparable to the approximate ASM fitting method described above.

To compare how well the ASMs and SSMs model the *Prince* data, we fit both models to the data for a range of dimensions q. Specifically, we found the subfamily in $F_A^{p,q}$ for each q from 0 to 112 minimizing (8), and found the subfamily in $F_S^{p,q}$ for q in the same range using the one-step approximation to the minimization of (14). Note that since there are $K = 113$ data points there will be subfamilies of dimension 112 in both $F_A^{p,q}$ and $F_S^{p,q}$ which will fit these data exactly.

Figures 3 and 4 compare the performances of the two classes of models. Since we have used different fitting criteria for each of the classes, we have two residual distance criteria with which to assess goodness of fit. Figure 3 plots the residual sum of squared geodesic distances (14) against q, for both models. For the ASMs this is calculated by putting $\tilde{\theta}_i = \sqrt{\tilde{\pi}_i}$ in (14), where the $\tilde{\pi}_i$ minimize (8).

The SSMs fit the data better according to this criterion, for any q. This is not surprising, since they were fit by minimizing an approximation to this criterion, whereas the ASMs were fit using the criterion (8). What is more interesting is that for most of the range of q the SSMs fit the data better according to criterion (8) too. This is shown in Figure 4, which plots (8) against q for the two models. Here we have calculated the residual distance for the SSMs by putting $\tilde{\pi}_i = \tilde{\theta}_i^2$ in (8), where the $\tilde{\theta}_i$ minimize (14). Note that this is the case even though the algorithm we have used to fit the SSMs minimizes (14) only approximately, but the ASM algorithm minimizes (8) exactly.

Of course, when comparing the performance of models using goodness of fit criteria, we have to take account of the relative sizes of the models. We show in Section 4 that fitting a (p, q)-ASM involves estimation of

$$\mathrm{df}_a = \mathrm{df}_a(p, q, K) = Kq + (q + 1)(p - q) \tag{15}$$

parameters. In Section 7 we show that fitting a (p, q)-SSM involves estimation of slightly more

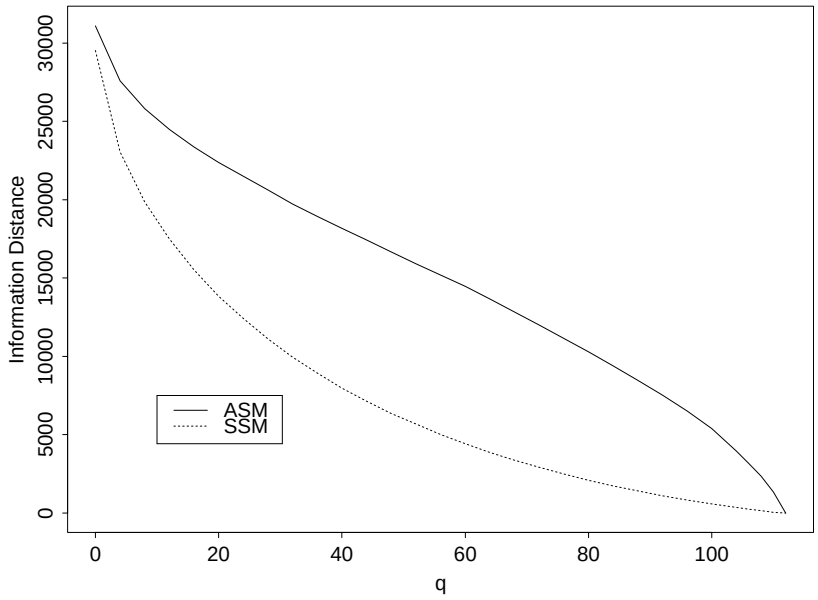


Figure 3: Residual squared geodesic distances (14) for SSMs and ASMs fit over a range of subfamily dimensions.

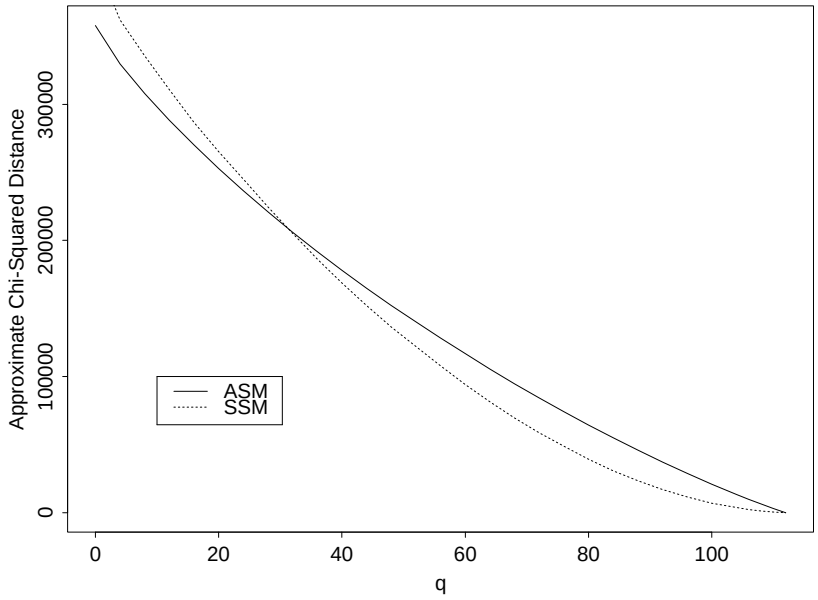


Figure 4: Residual approximate chi-squared distances (8) for SSMs and ASMs fit over a range of subfamily dimensions.

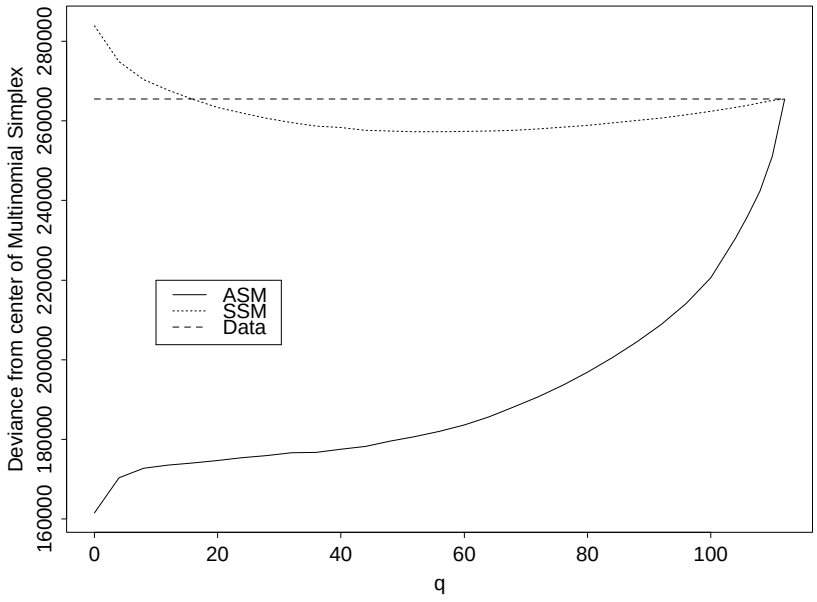


Figure 5: A measure of sparsity of the fits: deviances from fitted probability vectors to the center of the simplex, for the SSMs and ASMs of Figures 3 and 4.

parameters,

$$\mathrm{df}_s = \mathrm{df}_s(p, q, K) = Kq + (q+2)(p-q). \tag{16}$$

For large p and q this difference is small. For the *Prince* data, with $q = 60$ for example, we have df_a =287,319 and df_s =291,918, a difference of less than 2%. For this dimension of subfamily the SSM fit is 19% better by criterion (8) and 69% better by criterion (14). Since the methods we have used to fit the ASM and SSM models here are both only approximations to maximum likelihood methods, we cannot make precise asymptotic statements about the relative sizes of the models. But the improved performance of the SSMs over the ASMs seems far bigger than can be explained by the relatively small increase in model complexity.

We have suggested that the SSMs do better in this example because of the sparseness of the data, and the ability of the spherical subfamilies to curve around the edges of the probability simplex. We can test this explanation by comparing the sparsity of the ASM and SSM fits as follows. A simple measure of sparsity of a probability vector $\pi \in E_p$ is the deviance from that vector to the center of E_p, $\pi_c = (1, 1, \ldots, 1)/(p+1)$. From (6) this is

$$\mathrm{Dev}(\pi, \pi_c) = 2\log(p+1) - 2H(\pi), \tag{17}$$

where $H(\pi)$ is the entropy in π,

$$H(\pi) = -\sum_{j=1}^{p+1} \pi^j \log(\pi^j). \tag{18}$$

As a measure of sparsity of the probability vectors $\tilde{\pi}_1, \ldots, \tilde{\pi}_K$ fit by a model, we will use the weighted average of these deviances,

$$\sum_{i=1}^{K} n_i \mathrm{Dev}(\tilde{\pi}_i, \pi_c). \tag{19}$$

Figure 5 plots (19) for the ASMs and SSMs of the previous two figures. The horizontal line in the figure is the sparsity of the data. The SSMs achieve a level of sparsity much closer to the data, at much lower subfamily dimensions q, than do the ASMs. The ASM fits are much "smoother" than the data over most dimensions q, and their sparsity is only pulled up to match the data right at the highest levels of q.

Note
Model (1) may seem rather oversimplified for such a complex structure as document text, but actually works rather well. If each π_i^j is small, as is the case in this type of data, the counts x_i^j are approximately Poisson, with mean $n_i \pi_i^j$. Mosteller and Wallace (1984) test various models for word counts in text. They conclude that such counts are too overdispersed in the large documents they examine to support a Poisson model with a single mean throughout the document. But they find that an adequate model is one which breaks the text into short blocks with separate Poisson means for each block. This is, in effect, our model here. They look too at the correlations between the x_i^j for different j, concluding that the covariance structure of a multinomial model appears adequate.

3 Example: Compositional Data

Aitchison (1983) describes a method of principal component analysis for compositional data. He illustrates this method using data from Thompson et al. (1972) on chemical analyses of lavas from the Isle of Skye. Tables 1 and 2 of that paper list the percentages of 11 chemical components of 72 rock samples. For illustration, one such sample is

SiO_2	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	TiO_2	P_2O_5	MnO
46.47	13.94	10.99	12.80	9.77	2.10	0.40	1.53	0.11	0.18

The percentages in the samples sum to 100.

Instead of analyzing the data directly we will, for simplicity, follow Aitchison and restrict our attention to a summary of the data commonly used by geologists, the so-called AFM diagram. Writing A=$Na_2O + K_2O$ (the alkali content), F=Fe_2O_3 and M=MgO, let $\hat{\pi}_i = (\hat{\pi}_i^1, \hat{\pi}_i^2, \hat{\pi}_i^3)$,

$i = 1, \ldots, K = 72$, where $\hat{\pi}_i^1$, $\hat{\pi}_i^2$, and $\hat{\pi}_i^3$ are the proportions in the ith sample of A, F and M respectively, out of the total A+F+M. Each $\hat{\pi}_i$ lies in E_2, the trinomial simplex defined in (3).

Figure 6 shows the data points $\hat{\pi}_1, \ldots, \hat{\pi}_K$ plotted on E_2 in barycentric co-ordinates. The data follow an approximate curve through the multinomial simplex. Geologists classify rock samples by their relative positions along such curves, estimated from the data.

Recognising that a $(2,1)$-ASM (a straight line fit through this simplex) would be a poor model for these data, Aitchison proposed an alternative. The dashed line in the figure is a curve through the data estimated using Aitchison's method. It is the first principal axis from a principal component analysis of the log-transformed composition vectors. The solid line is the $(2,1)$-SSM estimate obtained using the iterative algorithm described in Section 8. The dotted line is another $(2,1)$-SSM estimate obtained using the one-step approximation to the previous estimate.

The data appear to follow a more curved path than is modeled by the Aitchison method. (See Note 1 below.) The SSMs look better. The same data are shown in Figure 7. Using the spherical parametrization with $\hat{\theta}_i = \theta(\hat{\pi}_i)$ and $\tilde{\theta}_i = \theta(\tilde{\pi}_i)$ as before, the x-axis in this figure is the angle of each fitted value $\tilde{\theta}_i$ around the fitted spherical subfamily, a circle in this parametrization. The (signed) residual distances $d(\hat{\theta}_i, \tilde{\theta}_i)$ for each point are plotted on the y-axis.

The parameterization of fitted values by angle as in Figure 7 is particularly convenient because, with a multinomial-like probability model for the data as described below, the parametrization is *variance-stabilizing*. That is, if we condition on the choice of fitted subfamily, the asymptotic standard errors of the positions of the points along the x-axis of the figure are all the same size. A 95% confidence interval based on this constant standard error is shown in the figure. Details of these calculations are in Section 5.

There is no standard probability model for compositional data. The summation constraints on the data vectors suggest though that the variance structure of these vectors could be similar to that of multinomial vectors. We will use a multinomial model for the data, putting

$$\hat{\pi}_i \sim \text{Multinomial}_p(n, \pi_i)/n, \quad i = 1, \ldots, K. \tag{20}$$

We will treat n as an unknown parameter to be estimated from the data. For the theory in the later sections of this paper to be applicable to these data we really only require that the π_i and $\hat{\pi}_i$ are in E_p, and that the $\hat{\pi}_i$ have first two moments

$$\text{E}[\hat{\pi}_i] = \pi_i, \quad \text{Var}[\hat{\pi}_i] = \sigma^2[\text{Diag}(\pi_i) - \pi_i \pi_i^T], \quad i = 1, \ldots, K, \tag{21}$$

for unknown σ^2. These conditions are satisfied by (20) with $\sigma^2 = 1/n$. We could, however, simply assume a quasi-likelihood (McCullagh and Nelder, 1989) which satisfies (21).

Since the SSM is fit by minimizing (14), a natural estimator of σ^2 in (21) uses the residual from this fit:

$$\tilde{\sigma}^2 = \frac{1}{K(p-q)} \sum_{i=1}^{K} d(\hat{\theta}_i, \tilde{\theta}_i)^2. \tag{22}$$

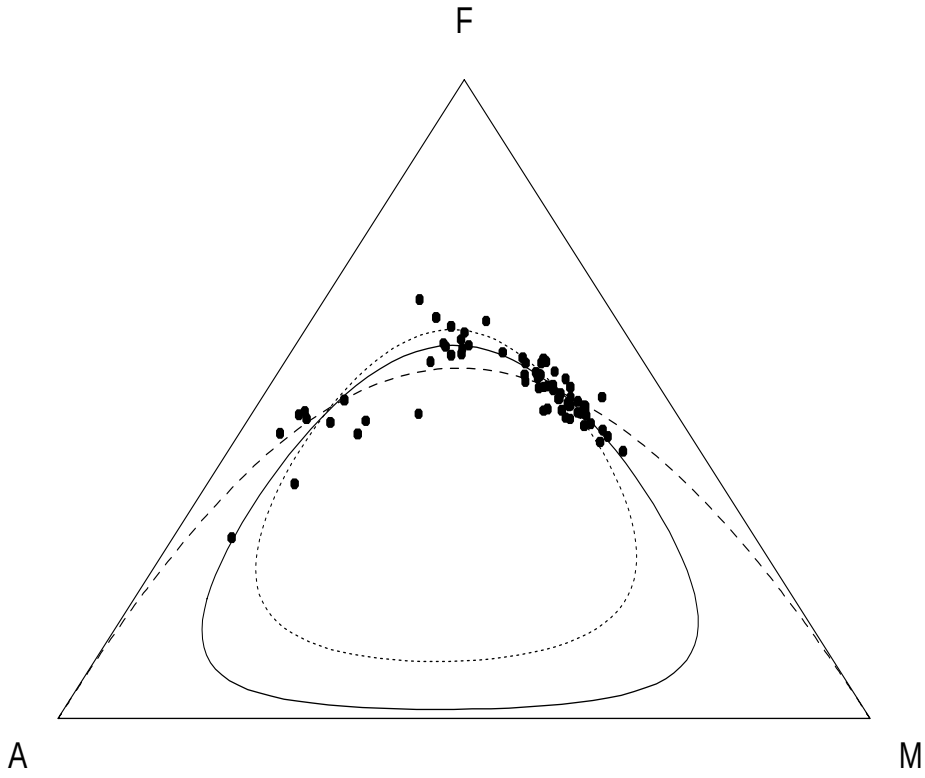


Figure 6: Three one-dimensional subfamilies fit to the Skye lava data of Section 3, plotted in barycentric co-ordinates. The solid and dotted curves are $(2,1)$-SSM estimates, using the iterative fitting algorithm and the one-step approximation to this algorithm respectively. The dashed curve is Aitchison's first principal axis.

Conditioning again on our choice of subfamily $\mathcal{F}$, this estimator is asymptotically unbiassed for σ^2, as is shown in Section 9.2. For these data we have $K(p-q) = 72$, and

$$\tilde{\sigma} = 0.060. \tag{23}$$

This corresponds to an effective n in (20) of 278.

Since $\mathcal{F}$ is chosen from the data, though, this estimate is too small. Heuristically, to take into account the estimation of the parameters describing $\mathcal{F}$, we should replace $K(p-q)$ in (22) by $Kp - \mathrm{df}_s(2,1,72)$. From (16) this is 69. In this case we obtain only a slightly larger estimate of σ^2, 0.062, and an effective n of 260.

The remaining sections develop the theory behind the results presented in Section 2 and this section.

Note 1

Only a subset of the data fitted here were used in Aitchison (1983): those in Table 2 of Thompson et al. (1972). This subset does not exhibit as much curvature as the full set, and in that case Aitchison's method is visually more adequate than it is here.

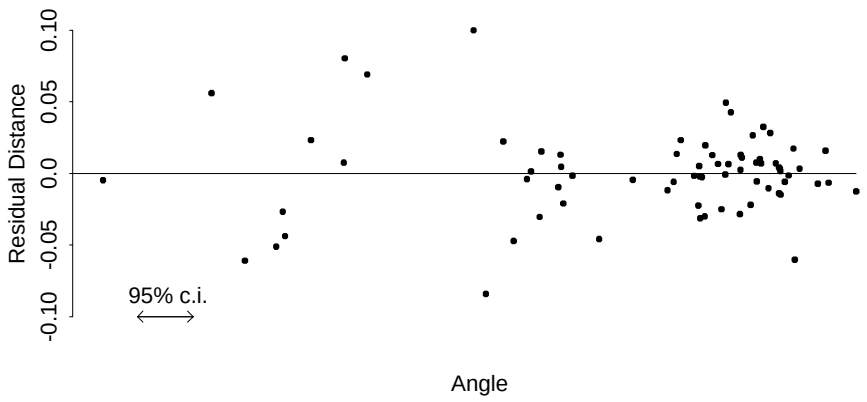


Figure 7: The Skye lava data plotted by angle along the fitted spherical subfamily.

Note 2

Aitchison's log transform cannot be performed on data vectors with zero entries. There are no zeros in the data in this example, and adding a small constant to all zero entries could, in any case, alleviate that problem. As Anderson (1982) notes, though, the choice of this arbitrary constant has a large effect on the resulting fit. This is a problem particularly if the method is used for multinomial data, not compositional data, in which zeros are more prevalent. The sparse data in Section 2 is an extreme example.

4 Affine Subspace Models for Multinomial Data

We will use the notation $G(p,q)$ to refer to the collection of q-dimensional *vector* subspaces in $\mathbb{R}^p$, and $AG(p,q)$ to refer to the collection of q-dimensional *affine* subspaces in $\mathbb{R}^p$. A subspace $A \in AG(p,q)$ may be parameterized as

$$A = \{\alpha + \Lambda\gamma : \gamma \in \mathbb{R}^q\}, \tag{24}$$

where α is a p-vector and Λ is a $p \times q$ matrix, satisfying $\alpha^T\Lambda = 0$ and $\Lambda^T\Lambda = I$. If $A \in G(p,q)$ then $\alpha = 0$.

Suppose we are given points $y_1, \ldots, y_K \in \mathbb{R}^p$, and non-negative weights $w_1, \ldots, w_K$. The Euclidean projection $\tilde{\mu}_i$ of a y_i onto an $A \in AG(p,q)$ is

$$\tilde{\mu}_i = \tilde{\mu}(A, y_i) = \alpha + \Lambda\Lambda^T y_i. \tag{25}$$

There is a well-known procedure for finding the $\tilde{A} \in AG(p,q)$ which minimizes the weighted sum

of squared Euclidean distances between the y_i and the $\tilde{\mu}_i$,

$$\sum_{i=1}^{K} w_i \|y_i - \tilde{\mu}_i\|^2, \tag{26}$$

which, for convenience, we will restate here in our notation. From the weighted mean of the y_i,

$$\bar{y} = \sum_i w_i y_i / \sum_i w_i, \tag{27}$$

construct the centered matrix

$$Y = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_K - \bar{y} \end{bmatrix}. \tag{28}$$

Calculate the SVD (see Golub and Van Loan, 1983),

$$\mathrm{Diag}(w_1, \ldots, w_K)^{\frac{1}{2}} Y = U \mathrm{Diag}(d_1, \ldots, d_p) V^T. \tag{29}$$

Here U is $K \times p$, V is $p \times p$, $U^T U = V^T V = I_p$. The α and Λ parameterizing $\tilde{A}$ are obtained from

$$V = [\Lambda : \Lambda^*_{p \times (p-q)}], \tag{30}$$

$$\alpha = (I - \Lambda \Lambda^T) \bar{y}. \tag{31}$$

There is a straightforward statistical interpretation of the above procedure. If we fix all $w_i = 1$, and assume that the y_i are independent with

$$y_i \sim N_p(\mu_i, \sigma^2 I), \quad i = 1, \ldots, K, \tag{32}$$

all μ_i in an unknown $A \in AG(p, q)$ and σ^2 known or unknown, then the $\tilde{\mu}_i = \tilde{\mu}(\tilde{A}, y_i)$ are the MLEs of the μ_i. This is because the squared Euclidean distances in (26) are proportional to the multivariate normal deviances.

The (p, q)-ASM model restricts the π_i in (1) to a subfamily in the collection

$$F_A^{p,q} = \{E_p \cap A : A \in AG(p, q),\ E_p \cap A \neq 0\}. \tag{33}$$

Fitted probability vectors $\tilde{\pi}_i$ minimizing the approximate chi-squared criterion (8) can be found simply by setting

$$y_i = \mathrm{Diag}(\hat{\pi}_0^{-\frac{1}{2}}) \hat{\pi}_i, \quad \text{and} \quad w_i = n_i, \quad i = 1, \ldots, K, \tag{34}$$

obtaining $\tilde{\mu}_i$ using the procedure above. Then put

$$\tilde{\pi}_i = \mathrm{Diag}(\hat{\pi}_0^{\frac{1}{2}}) \tilde{\mu}_i, \quad i = 1, \ldots, K. \tag{35}$$

This procedure does not guarentee that all $\tilde{\pi}_i \in E_p$. This is occasionally not the case, particularly for small q. If some $\tilde{\pi}_i^j < 0$ for some fit in the example of Section 2, we simply set $\tilde{\pi}_i^j$ to zero and renormalized $\tilde{\pi}_i$, so violating the affine restriction in those few cases. A similar correction must be made to certain SSM fits, as is explained in Section 7.

The (p,q)-ASM models are very similar, but not identical, to canonical analysis models. Put

$$\hat{p}_i = \frac{n_i \hat{\pi}_i}{\sum_i n_i}, \quad \tilde{p}_i = \frac{n_i \hat{\pi}_i}{\sum_i n_i}, \quad i = 1, \ldots, K. \tag{36}$$

Then $\hat{P} = [\hat{p}_1 \ldots \hat{p}_K]^T$ and $\tilde{P} = [\tilde{p}_1 \ldots \tilde{p}_K]^T$ are matrices of probabilities summing to 1. The affine restriction on the $\tilde{\pi}_i$ in a (p,q)-ASM is equivalent to a restriction of $\tilde{P}$ to rank $q+1$, subject to the row sums of the matrix being equal to those of $\hat{P}$. The canonical analysis models (as presented, for example, in Gilula and Haberman, 1986), also restrict the rank of $\tilde{P}$, but subject to both the row and column sums of the matrix being equal to those of $\hat{P}$.

The ASMs, through model (1), explicitly condition on the row sums of $\tilde{P}$. The canonical analysis models implicitly condition on both the row and column sums of $\hat{P}$ since any fit $\tilde{P}$ matches these sums.

Correspondence analysis fits a canonical analysis model to the matrix $\tilde{P}$ by minimizing the fitting criterion

$$\sum_{i=1}^{K} \sum_{j=1}^{p+1} \frac{(\hat{\pi}_i^j - \tilde{\pi}_i^j)^2}{n_i \hat{\pi}_0^j}, \tag{37}$$

using an SVD and a similar transformation to (34)–(35). If all the n_i are equal then, comparing (37) to (8), the correspondence analysis fit and the approximate chi-square fit we have used here are equivalent. This is the case for the *Prince* data.

The formula for the number of parameters fit by a (p,q)-ASM model, equation (15) of Section 2, can be obtained in two ways. The first is through differential geometry: $AG(p,q)$ may be endowed with the manifold structure of the Affine Grassmann manifold, which is of dimension

$$\dim\, AG(p,q) = (p+1)(p-q). \tag{38}$$

(See Okubo, 1987.) Once the subspace has been specified, a further Kq parameters place the $\tilde{\pi}_i$ on the subspace.

Alternatively, one can just count parameters. The parametrization (24) for $A \in AG(p,q)$ requires $p + pq$ parameters in total. The constraints $\alpha^T \Lambda = 0$ and $\Lambda^T \Lambda = I$ amount to $q + q(q+1)/2$ restrictions on these parameters. Also, this parametrization of A is only unique up to rotations of the columns of Λ:

$$A = \{\alpha + \Lambda M \gamma : \gamma \in \mathbb{R}^q\} \tag{39}$$

for any $q \times q$ orthogonal matrix M. A further $q(q-1)/2$ restrictions are required to fix a rotation, giving a total of $q^2 + q$ restrictions. Subtracting this from $p + pq$ gives the result.

Formula (15) is equivalent to that derived by Good (1965), p. 64, in the context of canonical models.

5 The Spherical Geometry

There is a well-known variance-stabilizing transformation for the binomial family. Suppose $\hat{\pi} \sim \text{Multinomial}_2(n, \pi_0)/n$, and

$$\hat{v} = \arccos(\sqrt{\hat{\pi}^1}). \tag{40}$$

Then a Taylor expansion of the above gives $\text{Var}(\hat{v}) \approx 1/4n$ for large n, an expression which does not depend on π_0. This result can be generalized to p-dimensional multinomial families. Suppose $\hat{\pi} \sim \text{Multinomial}_p(n, \pi_0)/n$. Let $\hat{\theta} = \theta(\hat{\pi}) \in S_+^p$. Since $\text{Var}[\hat{\pi}] = (\text{Diag}(\pi_0) - \pi_0\pi_0^T)/n$, for large n,

$$\text{Var}[\hat{\theta}] \approx \frac{1}{4n}(I - \theta_0\theta_0^T), \tag{41}$$

where $\theta_0 = \theta(\pi_0)$. Suppose c is a curve in S_p^+ parameterized by a distance function on S_+^p which is a multiple ρ of arclength. Then the length of the velocity vector along this curve is

$$\|\dot{c}(v)\| = \frac{1}{\rho}, \tag{42}$$

a constant, for all v in the domain of c. If $\theta_0 = c(v_0)$ for some v_0, then $\hat{v}$, the MLE of v_0, has variance

$$\text{Var}[\hat{v}] \approx \left[\, \dot{c}(\hat{v})^T \,\text{Var}[c(\hat{v})]^{-1}\, \dot{c}(\hat{v}) \,\right]^{-1}. \tag{43}$$

Since $\dot{c}(v)^T c(v) = 0$, and from the identity $(I - \theta\theta^T)^{-1} = (I + \theta\theta^T/(1 - \|\theta\|^2))$, we get

$$\text{Var}[\hat{v}] \approx \frac{\rho^2}{4n}. \tag{44}$$

This reduces to the binomial result for $p = 2$, $\sigma^2 = 1/n$, $c(v) = (\cos v, \sin v)$, $0 \le v \le \pi/2$, and $\rho = 1$.

Taking $n = 1$, (44) is the inverse of the Fisher information for v_0 in $\hat{v}$. The information distance for the multinomial family $\mathcal{M}_p$ is the multiple ρ of arclength distance which makes this information equal to 1. Setting $\rho = 2$, this is (13).

Define $\tilde{v}$ to be the minimum information distance estimate of v_0, that is, the v minimizing $d(\hat{\theta}, c(v))$ over v in the domain of c. By results (ii) of Section 9.2 $\tilde{v}$ has the same asymptotic variance (44) as $\hat{v}$. This result will be used in the next section.

The geometry of S_+^p, with the information distance as metric, is called the *information geometry* of $\mathcal{M}_p$. A rigorous development of the theory of information geometry in general statistical families is presented in Kass and Vos (1997).

6 Spherical Subfamilies

Any q-dimensional subsphere of S_p is an intersection of S_p with a $(q+1)$-dimensional affine subspace of the surrounding space $\mathbb{R}^{p+1}$. That is, it is of the form

$$\mathcal{F}_S(A) = S_p^+ \cap A, \tag{45}$$

for $A \in AG(p+1, q+1)$. We therefore define, for $p \geq q \geq 0$, the (p,q)-spherical subfamilies to be the collection

$$F_S^{p,q} = \{\mathcal{F}_S(A) : A \in AG(p+1, q+1),\ \mathcal{F}_S(A) \neq \emptyset\}. \tag{46}$$

Fix $A = \{\alpha + \Lambda\gamma : \gamma \in \mathbb{R}^{q+1}\}$ so that $\mathcal{F}_S(A) \in F_S^{p,q}$. As before, we require $\alpha^T\Lambda = 0$ and $\Lambda^T\Lambda = I$. The radius of this subsphere is any $\|\gamma\|$ with $\alpha + \Lambda\gamma \in S_p$. This is

$$r = r(A) = \sqrt{1 - \|\alpha\|^2}, \tag{47}$$

so $0 < r \leq 1$. For any $\theta \in S_p^+$. we can calculate the point $\tilde{\theta} = \tilde{\theta}(A, \theta)$ on $\mathcal{F}_S(A)$ with minimum geodesic distance $\arccos(\tilde{\theta}^T\theta)$ to θ. Write $\tilde{\theta} = \alpha + \Lambda\tilde{\gamma}$. Since arccos is monotone decreasing, $\tilde{\gamma}$ maximizes

$$(\alpha + \Lambda\gamma)^T\theta = \alpha^T\theta + \gamma^T\Lambda^T\theta \tag{48}$$

over all $\gamma \in \mathbb{R}^{q+1}$ subject to $\|\gamma\| = r$. Clearly, unless $\|\Lambda^T\theta\| = 0$, $\tilde{\gamma} = r\Lambda^T\theta/\|\Lambda^T\theta\|$. So we have

$$\tilde{\theta}(A, \theta) = \alpha + \frac{r\Lambda\Lambda^T\theta}{\|\Lambda^T\theta\|}, \quad \|\Lambda^T\theta\| \neq 0. \tag{49}$$

The points θ with $\|\Lambda^T\theta\| = 0$ are of the form $S_p^+ \cap G$ for some $G \in G(p+1, p-q-1)$, and are equidistant to all points in $\mathcal{F}_S(A)$. Since these points will occur with probability zero under realistic probability models we will not consider this case further.

The simplest spherical subfamilies are those of one dimension, which we will call *circular subfamilies*. These are of the form $\mathcal{F}_S(A)$, $A \in AG(p+1, 2)$.

A convenient parameterization for a circular subfamily is the angle ϕ around the circle. The curve

$$c(\phi) = \alpha + \Lambda(r(A)\cos\phi, r(A)\sin\phi)^T, \quad 0 \leq \phi < 2\pi, \tag{50}$$

traces around the subfamily. Since $\|\partial c/\partial\phi\| = r(A)$, by the results in Section 5 this parameterization is variance-stabilizing.

The subfamily fit to the lava data in Section 3 and shown as the solid curve in Figure 6 is a circular subfamily with radius $r = 0.37$. Let $\tilde{\phi}_i$, $i = 1, \ldots, K$, be the K fitted lava compositions in this parameterization, so $\tilde{\theta}_i = c(\tilde{\phi}_i)$. Conditioning on the choice of subfamily, from (23), (42), and (44), with $\rho = 1/r(A)$, the $\tilde{\phi}_i$ have approximate standard error

$$se(\hat{\phi}_i) = \frac{\tilde{\sigma}}{2r} = 0.081. \tag{51}$$

Figure 7 plots $s_i d(\hat{\theta}_i, \tilde{\theta}_i)$ against $\tilde{\phi}_i$, $i = 1, \ldots, K$, where s_i is $+1$ or -1 depending on whether $\hat{\theta}_i$ lies on the inside or outside of the subfamily. That is,

$$s_i = \begin{cases} +1\text{: } \|\hat{\theta}_i - \alpha\| \geq r, \\ -1\text{: } \|\hat{\theta}_i - \alpha\| < r. \end{cases} \tag{52}$$

Included in the figure is a line showing the length of the asymptotic two-sided 95% confidence interval for all the $\tilde{\phi}_i$, $1.96 \times 2 \times se(\hat{\phi}_i) = 0.32$.

In this section we have discussed estimation of points on a fixed spherical subfamily $\mathcal{F}_S(A)$. Now we turn to the problem of estimating the subfamily itself from the data.

7 Spherical Subfamily Models

Let $\hat{\pi}_1, \ldots, \hat{\pi}_K$ be independent MLEs of multinomial data as in (1)–(2). Changing to the spherical parameterization, with $\hat{\theta}_i = \sqrt{\hat{\pi}_i}$ and $\theta_i = \sqrt{\pi_i}$, the (p,q)-SSM for these data restricts the θ_i to lie in some $\mathcal{F}_S(A) \in F_S^{p,q}$, defined in (45)–(46). Let $\tilde{A}$ minimize

$$\text{geodist}(A) = \sum_{i=1}^{K} n_i \arccos(\hat{\theta}_i^T \tilde{\theta}(A, \hat{\theta}_i))^2 \tag{53}$$

over $A \in AG(p+1, q+1)$. The function $\tilde{\theta}$ is defined in (49). We estimate the θ_i by $\tilde{\theta}_i = \tilde{\theta}(\tilde{A}, \hat{\theta}_i)$, $i = 1, \ldots, K$.

The derivation of formula (16) for the number of parameters required to specify a (p,q)-SSM is similar to the derivation of (15) in Section 4. From (38) $(q+2)(p-q)$ parameters are required to fix an $A \in AG(p+1, q+1)$, and Kq parameters to place the $\tilde{\pi}_i$ on the spherical subfamily defined by A.

An iterative algorithm to minimize (53) will be presented in the next section. The algorithm is reasonably complex, but there is a good, simple first approximation: minimize (26) with $y_i = \hat{\theta}_i$ and $w_i = 1/\sigma_i^2$ using (27)–(31) to find $\tilde{A}$ and the $\tilde{\mu}_i$. Then set $\tilde{\mathcal{F}} = \mathcal{F}_S(\tilde{A})$, and $\tilde{\theta}_i = \tilde{\theta}(\tilde{A}, \tilde{\mu}_i)$. This was the approximation used to fit the SSMs in Section 2, since for such a large dataset the iterative procedure was computationally prohibitive.

The approximation also provides a good initial guess at the minimizing subfamily for use by the algorithm below. This first approximation to the circular subfamily in Section 3 is shown as the dotted curve in Figure 6. The solid curve was found by iterating from this initial value. In this case the algorithm converged acceptably after just 3 iterations.

Note that similarly to the approximate ASM fit of Section 4, there is no guarentee that the approximate SSM fit here will result in all $\tilde{\theta}_i \in S_p^+$. In practice, if some $\tilde{\theta}_i^j < 0$ for a fitted value $\tilde{\theta}_i$, we simply set $\tilde{\theta}_i^j = 0$ before transforming back to $\tilde{\pi}_i = \tilde{\theta}_i^2$, and then renormalize the elements of $\tilde{\pi}_i$ to sum to 1.

8 The Fitting Algorithm

Suppose we are given $\hat{\theta}_i \in S_+^p$, $i = 1, \ldots, K$, an approximate minimizing subfamily $\mathcal{F}_S(A_0)$ such as that described in the previous section, and points $\tilde{\theta}_i = \tilde{\theta}(A_0, \hat{\theta}_i)$ on A_0. We wish to use A_0 to find

an A_1 such that

$$\text{geodist}(A_1) < \text{geodist}(A_0). \tag{54}$$

The idea of the algorithm is to construct an additional set of points $\theta'_i = \theta'(A_0, \hat{\theta}_i, \lambda)$ with associated weights $w_i = w(A_0, \hat{\theta}_i, \lambda)$ (the role of $\lambda > 0$ will be explained below) so that the A_1 minimizing

$$\text{eucdist}(A) = \sum_{i=1}^{K} w_i \|\theta'_i - \tilde{\mu}(A, \hat{\theta}_i)\|^2 \tag{55}$$

over $A \in AG(p+1, q+1)$ satisfies (54). Since (55) can be minimized using the method of Section 4, we can then use A_1 similarly to improve the fit further, and so on. The algorithm takes the form of an "iterative SVD".

The θ'_i will be constructed so that small changes in A_0, hence in the $\tilde{\theta}_i$, will result in the same change in eucdist(A_0) as in geodist(A_0). This requirement will be satisfied if, for each i, the following two conditions hold:

(i) The Euclidean projection of θ'_i onto A_0 is $\tilde{\theta}_i$. Formally,

$$\tilde{\mu}(A_0, \theta'_i) = \tilde{\theta}_i. \tag{56}$$

(ii) Let

$$D_i = \frac{\partial}{\partial \theta} n_i d(\hat{\theta}_i, \theta)\Big|_{\theta = \tilde{\theta}_i} \quad \text{and} \quad E_i = \frac{\partial}{\partial \theta} w_i \|\theta'_i - \theta\|^2\Big|_{\theta = \tilde{\theta}_i}, \tag{57}$$

gradients at $\tilde{\theta}_i$ of the ith terms of geodist(A_0) and eucdist(A_0) respectively. We require that the projections of D_i and E_i onto the tangent space of S^p at $\tilde{\theta}_i$ be equal.

In (ii) we require that the specified projection vectors of these gradients be matched, not the vectors themselves, because each $\tilde{\theta}_i$ is restricted to lie in S^p.

These two conditions do not, by themselves, determine θ'_i uniquely. We will add a third requirement which determines how far the θ'_i are placed from the $\tilde{\theta}_i$:

(iii) $\|\tilde{\theta}_i - \hat{\theta}_i\| = \lambda \|\tilde{\theta}_i - \theta'_i\|$.

For a fixed λ all θ'_i and w_i are determined uniquely. Their construction, and uniqueness, is shown in the Appendix.

At the beginning of each iteration we set λ to 1. If, once A_1 is calculated, (54) is not satisfied, we can reduce λ and recalculate all $\theta'_i = \theta'(A_0, \hat{\theta}_i, \lambda)$ and their associated weights. For small enough λ, A_1 can be made arbitrarily close to A_0. So each $\tilde{\theta}(A_1, \hat{\theta}_i)$ can be made close enough to $\tilde{\theta}(A_0, \hat{\theta}_i)$ that by conditions (i) and (ii), unless A_0 is at a local minimum, (54) will be true.

9 Appendix

9.1 Construction of θ_i' and w_i.

These two quantities can be constructed to satisfy (i)–(iii) of Section 8 as follows. For clarity we will put

$$\kappa = \frac{r}{\|\Lambda^T \hat{\theta}_i\|} \tag{58}$$

so that $\tilde{\theta}(A_0, \hat{\theta}_i) = \alpha + \kappa \Lambda \Lambda^T \hat{\theta}_i$.

First, from (57) we calculate

$$D_i = \frac{-2n_i \arccos(\hat{\theta}_i^T \tilde{\theta}_i)}{\sqrt{1 - (\hat{\theta}_i^T \tilde{\theta}_i)^2}} \hat{\theta}_i, \tag{59}$$

$$E_i = -2w_i'(\theta_i' - \tilde{\theta}_i). \tag{60}$$

To satisfy (ii) we need

$$(I - \tilde{\theta}_i \tilde{\theta}_i^T) D_i = (I - \tilde{\theta}_i \tilde{\theta}_i^T) E_i. \tag{61}$$

Some algebra shows that for the lengths of these two vectors to be equal we should set

$$w_i = \frac{n_i \arccos(\hat{\theta}_i^T \tilde{\theta}_i)}{(I - \tilde{\theta}_i \tilde{\theta}_i^T)\theta_i'}. \tag{62}$$

It is clear that these same two vectors will be collinear if and only if $\theta_i' \in \text{span}\{\tilde{\theta}_i, \hat{\theta}_i\}$. So write $\theta_i' = a\hat{\theta}_i + b\tilde{\theta}_i$, where $a, b \in \mathbb{R}$. Then (ii) is satisfied. For (i) we need

$$\tilde{\theta}_i = \tilde{\mu}(A_0, a\tilde{\theta}_i + b\hat{\theta}_i) = \alpha + (a\kappa + b)\Lambda\Lambda^T \hat{\theta}_i. \tag{63}$$

So put $a\kappa + b = \kappa$, and

$$\theta_i' = \theta'(A_0, \hat{\theta}_i) = a\tilde{\theta}_i + \kappa(1 - a)\hat{\theta}_i. \tag{64}$$

Finally, a can be chosen so that (iii) is satisfied. Since

$$\|\hat{\theta}_i - \tilde{\theta}_i\| = 2 - 2\hat{\theta}_i^T \tilde{\theta}_i, \quad \text{and} \tag{65}$$

$$\lambda\|\theta_i' - \tilde{\theta}_i\| = \lambda(a - 1)(1 - 2\kappa\hat{\theta}_i^T \tilde{\theta}_i + \kappa^2), \tag{66}$$

(iii) will be satisfied with

$$a = \frac{1}{\lambda}\left(1 + \frac{2 - 2\hat{\theta}_i^T \tilde{\theta}_i}{1 - 2\kappa\hat{\theta}_i^T \tilde{\theta}_i + \kappa^2}\right). \tag{67}$$

9.2 Inference for Minimum Information Distance Estimation

We state here two basic asymptotic results for minimum information distance estimation. We follow closely the theory and notation of Kass and Vos (1997), Section 7.6. All references below refer to that section.

Let $\mathcal{M}$ be an m-dimensional model space, $\mathcal{M}_0 \in \mathcal{M}$ a k-dimensional model subspace. Let $p_0 \in \mathcal{M}_0$, and $\hat{p}_n$ be the MLE of p_0 in $\mathcal{M}$ based on n independent samples. Let $\tilde{p}_n$ be the point in $\mathcal{M}_0$ of minimum information distance d to $\hat{p}_n$. Let θ and ϕ be parameterizations for $\mathcal{M}$ and $\mathcal{M}_0$ respectively, $\hat{\theta} = \theta(\hat{p}_n)$, $\tilde{\phi} = \phi(\tilde{p}_n)$, $\theta_0 = \theta(p_0)$, $\phi_0 = \phi(p_0)$. We assume $\hat{\theta} \Rightarrow N_m(\theta_0, i(\theta_0)^{-1})$, where $i(\theta_0)$ is the information matrix at θ_0. Then

(i) $\mathrm{n}d^2(\hat{\theta}, \tilde{\theta}) \Rightarrow \chi^2_{m-k}$,

(ii) $\tilde{\phi} \Rightarrow N(\phi_0, i(\phi_0)^{-1})$.

The first part is proved in Theorem 7.6.3. The second can be proved using results obtained in the proof of this theorem. Let $A = \exp^{-1}(\hat{p}_n)$, $B = \exp^{-1}\tilde{p}_n$, where exp is the exponential map on $\mathcal{M}$ at p_0. Let V_0 be the tangent space $T_{p_0}\mathcal{M}_0$, and define the V_0 projections $C = P_{V_0}A$ and $B' = P_{V_0}B$. Then

$$\|C\| \Rightarrow N_{V_0}(0, I_{V_0}) \tag{68}$$

with respect to the information metric, where I_{V_0} is the identity on V_0. Once we show $B = o(r)$ where $r = \|A\|$, the result will follow.

We have $BC < BB' + B'C$ ($BC = \|B - C\|$ etc), and BB' is $o(r)$, (7.6.6). By Pythagoras

$$B'C^2 = AB'^2 - AC^2 \tag{69}$$

$$= (AB' - AC)(AB' + AC). \tag{70}$$

By (7.6.4), (7.6.5), and the equation following (7.6.7), $AB' - AC$ is $o(r)$. From this and because $AC < \|A\|$, $AB' + AC$ is $O(r)$. So $B'C = o(r)$ and we are done.

For our purposes we take $\mathcal{M} = \mathcal{M}_p$, the p-dimensional multinomial family (4), and $\mathcal{M}_0$ some fixed $\mathcal{F}_S \in F_S^{p,q}$, so that $m = p$ and $k = q$. From (i), with $\sigma^2 = 1/n$, the asymptotic distribution of (22) is $\sigma\chi^2_{K(p-q)}/K(p-q)$, and $\tilde{\sigma}$ is unbiassed.

References

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, *70*(1), 57–65.

Anderson, J. A. (1982). Discussion of "The statistical analysis of compositional data". *Journal of the Royal Statistical Society, Ser. B*, *44*(2), 161–163.

Berry, M. W., Dumais, S. T., and O'Brien, G. (1995). Using linear algebra for intelligent infomation retrieval. *SIAM Review*, *37*(4), 573–595.

Gilula, Z., and Haberman, S. J. (1986). Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association*, *81*(395), 780–788.

Golub, G. H., and Van Loan, C. F. (1983). *Matrix Computations.* John Hopkins University Press.

Good, I. J. (1965). *The estimation of probabilities.* MIT Press.

Greenacre, M. J. (1984). *Theory and Application of Correspondence Analysis.* Academic Press.

Greenacre, M. J., and Hastie, T. J. (1987). The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association*, *82*, 437–447.

Hirschfield, H. O. (1935). A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society*, *31*, 520–524.

Kass, R. E., and Vos, P. W. (1997). *Geometrical Foundations of Asymptotic Inference.* Wiley.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall.

Mosteller, F., and Wallace, D. L. (1984). *Applied Bayesian and classical inference: The case of The Federalist papers* (2nd ed.). Springer-Verlag.

Okubo, T. (1987). *Differential Geometry.* M. Dekker.

Thompson, R. N., Esson, J., and Dunham, A. C. (1972). Major element chemical variation in the Eocene lavas of the Isle of Skye, Scotland. *Journal of Petrology.*