Application of Compositional Models for Glycan HILIC Data

Marie Galligan[13], Matthew P. Campbell[2], Radka Saldova[2], Pauline M. Rudd[2], Thomas Brendan Murphy[1]

[1]Department of Mathematical Sciences - University College Dublin, Ireland marie.galligan@ucd.ie
[2] National Institute for Bioprocessing Research and Training, Ireland
[3] Irish Research Council for Science, Engineering & Technology (IRCSET)

Glycoconjugates constitute a major class of biomolecules which include glycoproteins, glycosphingolipids and proteoglycans. The enzymatic process in which glycans (sugar chains) are linked to proteins or lipids is called *glycosylation*. Glycosylation is involved in many biological processes, both physiological and pathological, inlcuding host-pathogen interactions, tumour invasion, cell trafficking and signalling. Changes in glycan structure are thought be be at least partly responsible for the development of inflammation, infection, arteriosclerosis, immune defects and autoimmunity. Such changes have been observed in human diseases such as diabetes mellitus, rheumatoid arthritis and Alzheimer's Disease. Aberrant patterns of glycosylation are also a universal feature of cancer cells. The field of glycobiology thus shows great potential for the discovery of glycan biomarkers for disease diagnosis and prognosis.

Here we focus specifically on N-glycans, that is, glycans attached to protein molecules via a nitrogen atom. This class of glycans is the best characterized. High-throughput HILIC analysis is a well-established technique for the separation and quantification of N-linked glycans released from glycoproteins. HILIC analysis quantifies the N-glycan structures in serum via a chromatogram, which is subsequently standardized and integrated. The generated data for each sample is a set of relative HILIC peak areas and as a result, the data is compositional. To-date, most statistical analyses of these glycan data fail to account for their compositional nature.

We compare and contrast three compositional data models for the glycan HILIC data: the Dirichlet, Nested Dirichlet and Logistic Normal models, with the intention of providing tools for the statistical analysis of compositional data analysis in the glycobiology field. We use these three models for classification of disease/control cases in ovarian and lung cancer diagnosis applications. We discuss and compare these models in terms of their classification performance and goodness-of-fit.

1 Introduction

Glycobiology is the study of the structure, biology and biosynthesis of glycans. Glycans are chains of monosaccharides (simple sugar units) which are often linked to protein or lipid molecules, to form *glycoconjugates*.The biological roles of glycans are diverse and range from structural to mediatory. Taylor and Drickamer (2003) divide their functions into two main groups - intrinsic (such as providing structural components for cell walls, modifying stability and solubility properties of proteins) and extrinsic (the functions based on the recognition of glycans by other molecules). Glycosylation has been heavily implicated in human disease. Aberrant patterns of glycosylation have been shown to be a potential source of disease biomarkers, particularly in the diagnosis and prognosis of cancer and chronic inflammation (Arnold et al., 2008).

The search for N-glycan biomarkers is ongoing in the Glycobiology field. For example, Kyselova et al. (2008) identify several N-glycan structures which appear to differ in breast cancer patients, when compared with healthy control samples. Ercan et al. (2010) find aberrant patterns in the N-glycosylation of IgG glycoproteins from rheumatoid arthritis patients. The identification of valid glycan biomarkers could be of great value in disease diagnostics and therapeutics.

Royle et al. (2008) outline a high-throughput hydrophilic interaction liquid chromatography (HILIC) analysis for detailed quantitative analysis of serum N-glycans released from glycoproteins. The relative quantites of N-glycans are reported via a glycan profile similar to those in Figures 1 and 2. Each such HILIC chromatogram generates a set of peaks and their relative percentage areas. The nature of the data gives rise to a set of compositional vectors (one for each sample), such that each part of the composition represents the relative proportion of the area under the profile accounted for by a given peak.

To-date, most statistical analyses of these HILIC glycan profiles fail to account for the compositional nature of these data. Our objective is to establish an improved method for the statistical analysis of glycan HILIC data, for the identification and validation of glycan biomarkers. In this paper, we compare and contrast three compositional data models - the Dirichlet, the Nested Dirichlet and the Logistic Normal models. We fit each model to two glycan HILIC data sets, which are outlined in Section 2. A brief review of each model is included in Section 3, along with some information on model fitting strategies and the model comparison criteria used. We apply each of these models to the two data sets and use the fitted models for supervised classification of samples. The results are presented in Section 4, followed by a short discussion of the findings in Section 5. In Appendices A - C we include some additional information on parameter estimation for the Dirichlet distribution, the derivation of the Nested Dirichlet distribution and the techniques used for N-glycan HILIC analysis.

2 Data

2.1 Lung Cancer Data

Serum samples from preoperative patients diagnosed with lung cancer and cancer-free healthy volunteers were obtained from Fox Chase, Cancer Center, Philadelphia, USA under IRB approved protocols. Arnold et al. (2011) quantified the N-glycans in these serum samples, which consist of 100 lung cancer patients (20 from each of 5 stages - I, II, IIIA, IIIB, IV) and 84 age-matched controls (donors who did not have cancer). The analysis was carried out by HILIC floresence over using a 60-minute method. Further details on the analysis may be found in Arnold et al. (2011).

The resulting chromatogram from each sample was integrated over the set of 17 glycan peaks, resulting in a 17 part compositional vector for each observation. A typical profile from the analysis is shown in Figure 1.

For the purposes of statistical analyses, we re-group the data into 84 controls and 100 lung cancer patients.

Table 1: Details of the 184 patient samples used in the lung cancer study

	Disease Status	Cases
1	Controls	84
2	Stage 1	20
3	Stage 2	20
4	Stage 3A	20
5	Stage 3B	20
6	Stage 4	20

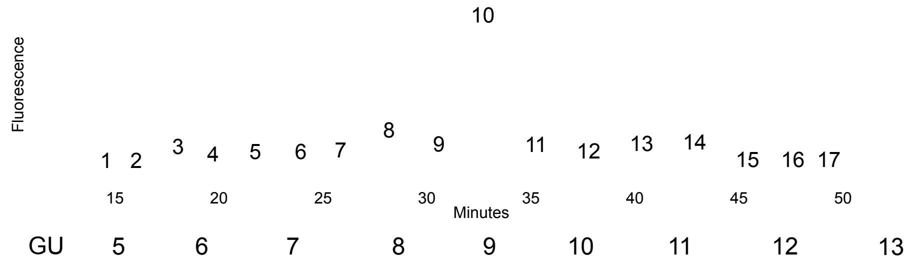

Figure 1: Typical HILIC chromatogram of N-glycans released from serum glycoproteins for the Lung Cancer Cohort (1hr. HILIC, integrated into 17 peaks). Each peak represents major N-glycan structures found in serum. The complete serum N-glycome is described in Royle et al. (2008). Further details on glycan structures contained under each peak may be found in Table 14

2.2 Ovarian Cancer Data

Our second data set consists of a set of 24 compositional variables for 63 observations, obtained from the ovarian cancer data study in Saldova et al. (2007). Venous blood samples were obtained from 63 serum samples taken from healthy controls, patients with benign gynecological conditions, borderline ovarian tumors, ovarian cancer, primary peritoneal carcinomatosis (PPC), endometrial cancer metastasized to the ovary and other gynecological cancers. These groups are quantified in Table 2. The samples were collected at St. James's University Hospital (Leeds, UK), following ethical approval and obtaining informed consent. After allowing the blood to clot for 30-60 min., serum was obtained by centrifugation at 2000xg for 10 minutes and stored at $-80\,^{\circ}\mathrm{C}$ until analysis. The N-glycans were analysed by HILIC fluorescence, using the 120 minute method, and the chromatograms were integrated into 24 glycan peaks, resulting in a set of 24 compositional variables for each sample. A typical chromatogram from the analysis is displayed in Figure 2.

Table 2: Details of the 63 patient samples used in the ovarian cancer study

	Disease Status	Cases
1	Controls	3
2	Benign	12
3	Borderline	6
4	Ovarian Cancer	13
5	Primary peritoneal carcinomatosis (PPC)	3
6	Other gynecological cancers metastasizing to ovary	5
7	Other gynecological cancers	21

We investigate two different re-groupings of the data for model fitting. In Section 4.2, we define two groups in the data. We merge the control and benign groups (1 and 2), as these are free from ovarian cancer and merge the ovarian cancer groups (4 and 6). We then use this to fit control vs.

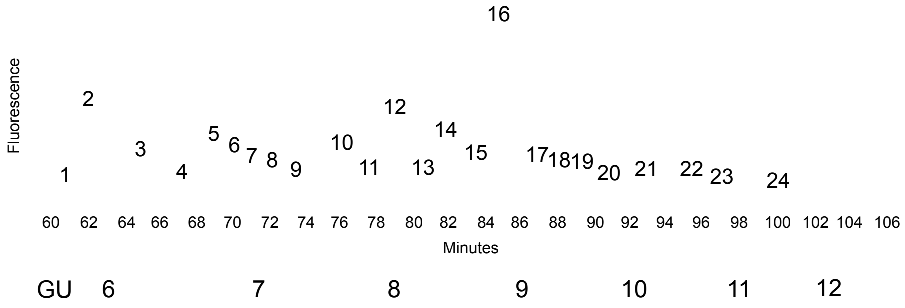

Figure 2: Typical HILIC chromatogram of N-glycans released from serum glycoproteins for the Ovarian Cancer Cohort (2hrs. HILIC, integrated into 24 peaks). Each peak represents major N-glycan structures found in serum. The complete serum N-glycome is described in Royle et al. (2008). Further details on glycan structures contained under each peak may be found in Table 15.

ovarian cancer models. We also define a three group model for this data set. In Section 4.3 we fit our models to just three subgroups of the data : the benign group (2), the ovarian cancer group (4) and the group with other gynecological cancers (7), as these groups would seem, intuitively, to have least overlap.

3 Methods

A data vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ consisting of strictly positive real numbers, which are constrained to sum to a fixed value, is known as a *composition*. The sample space of such data is a $(p-1)$-simplex embedded in $\mathbb{R}^p$ space, defined by

$$\mathbb{S}^{(p-1)} = \{\mathbf{x} = (x_1, x_2, \ldots, x_p) \in \mathbb{R}^p;\ x_1 \geq 0, \ldots, x_p \geq 0;\ \sum_{j=1}^{p} x_j = K\}$$

where K is an arbitrary constant. The data is typically scaled such that $K = 1$. Thus the sample space of the data is a subspace of $\mathbb{R}^p_+$, the positive orthant of p dimensional real space. This should be accounted for in the statistical analysis of the data.

Many methodologies for compositional data analysis have been discussed and tested in the literature, and two main parametric classes of models have emerged. One is the Dirichlet class of models, which model the data directly within the simplex, and the other is the Logistic Normal class of models. Taking this approach, the data is transformed to real space and the multivariate normal model is fitted to the transformed data. There are favourable and unfavourable arguments for both approaches to modeling the data.

The Dirichlet family has some very elegant mathematical properties and the models are easily interpretable in terms of the composition. Nevertheless, it has been well documented that the ordinary Dirichlet distribution is generally incapable of capturing some of the patterns of variance arising in real data, due in no small part to its strong implied independence structure. However, in an attempt to search for distributions on the simplex with less enforced independence, more flexible extensions of the Dirichlet distribution have also been considered with some success, such as the Generalized Dirichlet distribution, introduced by Connor and Mosimann (1969). We examine one such model called the Nested Dirichlet model (Null, 2008) described in Section 3.2. We fit both this and the ordinary Dirichlet model to the data sets outlined in Section 2 and compare the results. We also consider the second approach, that of modeling the data using the Logistic Normal model proposed by Aitchison and Shen (1980), which is outlined in Section 3.3; Aitchison (2003) describes many advantages of this model. The Logistic Normal class has a richer structure than the Dirichlet class and thus it has the ability to capture patterns of variability in real data which are beyond reach of the Dirichlet class. Importantly, since the Logistic Normal model is so closely related to the well-established multivariate normal class, we can easily utilize the wide range of tools which have been developed for this class.

We compare and contrast each of these models (the Dirichlet, the Nested Dirichlet and the Logistic Normal models) when applied to our data. We evaluate them on both their predictive abilities (using their cross-validated supervised classification performance) and their goodness-of-fit, measured by the Bayesian Information Criterion (BIC). Our objective is to make tools available for the statistical analysis of glycan HILIC data, which allows the three models to be fitted and a clear comparison made between them.

3.1 The Dirichlet Model

The Dirichlet probability density function for a vector of compositional components $\mathbf{x} = (x_1, x_2, \ldots, x_p) \in \mathbb{S}^p$, such that $x_l \geq 0$ for $l = 1, 2, \ldots, p$ and $\sum_{l=1}^{p} x_l = 1$ is written

$$f(\mathbf{x}; \mathbf{w}) = \frac{1}{B(\mathbf{w})} \prod_{l=1}^{p} x_l^{\omega_l - 1}$$

where $\mathbf{w} = (\omega_1, \omega_2, \ldots, \omega_p) \in R^p$, such that $\omega_i > 0$ for $i = 1, 2, \ldots, p$ and $\omega^* = \sum_{l=1}^{p} \omega_l$, is the Dirichlet parameter vector. $B(\mathbf{w})$ is the multinomial beta function defined as

$$B(\mathbf{w}) = \frac{\prod_{\mathbf{l=1}}^{\mathbf{p}} \mathbf{\Gamma}(\omega_\mathbf{l})}{\mathbf{\Gamma}(\omega^*)}$$

The Dirichlet distribution may be derived from a set of independent gamma random variables with the same scale parameter. Given a vector of p independently distributed Gamma random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$, each with scale parameter θ and with shape parameters $(\omega_1, \omega_2, \ldots, \omega_p)$, respectively, the Dirichlet distribution is directly recoverable through the change of variables $\mathbf{X} = \mathbf{Y} / \sum_{i=1}^{p} Y_i$. The transformed variables $\mathbf{X}$ are said to follow a Dirichlet distribution with parameters $(\omega_1, \omega_2, \ldots, \omega_p)$. This property gives the Dirichlet distribution a strong implied independence structure, making this model generally incapable of capturing the true variance structure of a set of real data. This said, the Dirichlet distribution also has some very elegant properties, such as the following agglommeration property: If

$$(x_1, x_2, \ldots, x_p) \sim \text{Dir}(\omega_1, \omega_2, \ldots, \omega_p)$$

then we may join any pair of these variables x_i and x_j and easily derive their new distribution as

$$(x_1, \ldots, (x_i + x_j), \ldots, x_p) \sim \text{Dir}(\omega_1, \ldots, (\omega_i + \omega_j), \ldots, \omega_p)$$

This may be useful for the glycan HILIC data. For example, the combined relative areas of particular peaks could be of interest as a disease marker, rather than, or as well as, the individual relative areas. This property is also very useful in the derivation of the Nested Dirichlet model outlined in Section 3.2.

3.1.1 Fitting a Dirichlet Distribution

To estimate the parameters for the distribution, we use the fixed point iteration method outlined by Minka (2000) to numerically maximize the log likelihood function. (Huang, 2005) compares four methods of estimating Dirichlet distributions and concludes that the fixed point iteration was the only one of the four that had held the inequality $\omega_l > 0$ for all l, without having to place a bound on the objective function.

To carry out the fixed point iteration, we begin with an initial guess for $\mathbf{w}$ and optimize the log liklihood with respect to $\mathbf{w}$, iterating over the following steps for iteration $t = 1, 2, \ldots$ until convergence

- Find a function $l_{\text{lower}}(\mathbf{w}^t|\mathbf{X})$ that places a lower bound on the log likelihood, $l(\mathbf{w}_t|\mathbf{X})$ which is tight at the current estimate $\mathbf{w}^{t-1}$ (a fixed quantity).
- Maximize the lower bound $l_{\text{lower}}(\mathbf{w}^t|\mathbf{X})$ with respect to $\mathbf{w}^t$ to update the parameter estimates.

This method ensures that the log likelihood increases at each step and reaches a maximum when the parameters converge.

Starting values for approximating MLEs for the Dirichlet

We use a variation on the Method of Moments, suggested by Ronning (1989), to find starting values for the MLE approximation. The first and second moments of the Dirichlet distribution are:

$$\mathbb{E}[X_k] \;=\; \frac{\omega_k}{\omega^*} \quad \text{and} \quad \mathbb{E}[X_k^2] \;=\; \mathbb{E}[X_k]\frac{1+\omega_k}{1+\omega^*}$$

and then variance of the Dirichlet may be written

$$\text{Var}[X_k] = \frac{\mathbb{E}[X_k](1-\mathbb{E}[X_k])}{1+\omega^*}$$

Re-arranging the above equation gives an expression for the sum of the parameters, in terms of the expectation and variance of the Dirichlet, which we estimate from our data.

$$\omega^* \;=\; \frac{\mathbb{E}[X_k](1-\mathbb{E}[X_k])}{\text{Var}[X_k]} - 1$$

However, rather than using just a single compositional variable X_k to estimate this sum, Ronning (1989) suggests that we average this quantity over $(p-1)$ variables, to give a more reliable estimate. Hence, the initialization we use for the fixed point iteration is

$$\log \widehat{\omega^*} \;=\; \frac{1}{p-1}\sum_{k=1}^{p-1} \log\left(\frac{\mathbb{E}[X_k](1-\mathbb{E}[X_k])}{\text{Var}[X_k]} - 1\right)$$

We then calculate an initial parameter estimate for each variable, using its expectation

$$\widehat{\omega_k} \;=\; \widehat{\omega^*}\mathbb{E}[X_k]$$

.

A Fixed Point Iteration to optimze MLEs for the Dirichlet

Given our starting values for w, we numerically maximize the Dirichlet likelihood function updating our parameters through iteration $t = 1, 2, \ldots$ until convergence of the model log likelihood. We describe the fixed point iteration method in detail in Section A. The parameter update at iteration t is

$$\omega_k^t = \Psi^{-1}\left(\Psi(\sum_{l=1}^{p}\omega_l^{t-1}) + \frac{1}{N}\sum_{i=1}^{n}\log x_{ik}\right)$$

Note that this update involves inverting the digamma function, which must also be approximated numerically. We use Newton's Method for the approximation. Letting $s = \Psi\ \sum_{l=1}^{p}\omega_l^{t-1}) + \frac{1}{N}\sum_{i=1}^{n}\log x_{ik}$, we initialize the inverse digamma function using the approximation

$$\omega_k^{init} = \Psi^{-1}(s) \approx \begin{cases} \exp(s) + 0.5 & \text{if } s \geq -2.22; \\ -\frac{1}{s+\gamma} & \text{if } s < -2.22 \end{cases}$$

and iterate over

$$\omega_k^{new} = \omega_k^{old} - \frac{\Psi(\omega_k^{old}) - s}{\Psi'(\omega_k^{old})}$$

until convergence.

3.2 The Nested Dirichlet Model

An extension of the Dirichlet distribution is the more flexible Nested Dirichlet model. This model was introduced in Null (2008) and Null (2009). The latter applies the Nested Dirichlet to model base player abiility. It is used as a conjugate prior for multinomial data, gathered on fourteen possible outcomes from a plate appearance. This results in a Nested Dirichlet posterior distribution for player ability.

The Nested Dirichlet distribution extends the Dirichlet distribution by arranging the variables into a tree structure. The original compositional variables are leaves in the tree and we introduce k new variables, or *nesting variables*, which are the internal nodes in the tree. Each nesting variable is evaluated as the sum of variables nested directly underneath. We fit this model directly to our compositional data sets and use it for classification of samples. The structure of the model is outlined further below.

Some Notation for the Nested Dirichlet Distribution

- p: the number of "original" variables (which are the leaves in the nesting tree)
- k: the number of nesting variables
- x_i such that $i \leq p$ is the ith variable in the original compositional vector, which has a corresponding Nested Dirichlet parameter ω_i
- $I_j = \{i : x_i \text{ is nested under } x_{p+j}; j = 1, 2, \ldots, k\}$, the set of indices for variables in nest j.
- x_{p+j} is the jth nesting variable, which has a corresponding parameter ω_{p+j}

- $\mathbf{W}_j = \{\omega_l : l \in I_j\}$ represents the set of Nested Dirichlet parameters for variables in the jth nest

- $\mathbf{W}_j^* = \sum_{l \in I_j} \omega_l$ the sum of parameters of variables in the jth nest

The density of the Nested Dirichlet distribution is derived from the product of the $k+1$ distributions of each nest and results in the pdf

$$f(\mathbf{x}; \boldsymbol{\omega}\ ) = \frac{\prod_{l=1}^{p+k} x_l^{\omega_l - 1}}{\prod_{s=0}^{k} B(\mathbf{W}_s) \prod_{j=1}^{k} x_{p+j}^{\mathbf{W}_j^* - 1}}$$

The probability density function for the Nested Dirichlet may be derived from a product of the Dirichlet distributions contained within the nests, as shown in appendix B

3.2.1 Fitting a Nested Dirichlet Distribution

For p variables, there are p^{p-2} possible nesting trees. Therefore, for large p, it would be computationally inefficient to search through all possible solutions. We use a heuristic search to find a 'good' nesting tree, by optimizing the BIC (Bayesian Information Criterion) defined in Section 3.4. The following forwards-searching algorithm seems to give good results:

1. Decide on a value for the maximum number of iterations (MAX_ITER) and maximum number of successive rejections (MAX_REJ)

2. Begin with a random initialzation of the nesting tree

3. At each iteration, propose one of the following moves:

   (a) **CREATE** Create a new nest by joining two variables

   (b) **ADD** Add variable i to an already existing nest j (if i and j are in the same nest)

   (c) **JOIN** Join two existing nests i and j (if i and j are in the same nest)

   (d) **BREAK** Propose to remove a subset of variables from an existing nest

4. Accept the proposed move if it increases the model BIC

5. Iterate over steps 2 and 3 until MAX_ITER iterations are carried out or MAX_REJ proposals are successively rejected

We run this algorithm over several initializations and choose the final nesting tree which has the largest BIC.

We apply this forwards search to the data sets outlined in and the results are shown in 4. For each data set, the algorithm was run over each of 100 random initializations of the nesting tree. We defined the algorithm as having converged if either

1. a maximum of 150 (MAX_ITER) iterations was reached.

2. 50 (MAX_REJ) successive proposed moves were rejected.

Parameter Estimation for the Nested Dirichlet

The distribution of each nest in the Nested Dirichlet model, conditional on its sum (i.e. conditional on the variable it is nested under), is independent of the variables outside that nest. Hence, to estimate the parameters for the Nested Dirichlet, we find use the method outlined in Section 3.1.1 to estimate the Dirichlet parameters for each nest. Thus, each node in our nesting tree will have a maximum likelihood parameter estimate, which is also its MLE for the Nested Dirichlet model.

3.3 The Logistic Normal Model

A class of distributions have been developed for modeling data on the simplex, derived from the normal distribution. Aitchison and Shen (1980) outline some properties and applications of this class of distributions. A compositional vector $\mathbf{x}$ with p components, such that the components sum to one, lies in the simplex space defined by

$$\mathbb{S}^d = \{(x_1, x_2, \ldots, x_d) \in \mathbb{R}^d_+;\ x_1 \geq 0, \ldots, x_d \geq 0;\ x_1 + x_2 + \ldots + x_d \leq 1\}$$

where $d = p - 1$ and $\mathbb{R}^d_+$ is the positive orthant of $\mathbb{R}^d$ space. Note that x_p is determined by the first d components, since $x_p = 1 - x_1 - \ldots - x_d$. Then any compositional vector $\mathbf{x} \in \mathbb{S}^d$ may be transformed to $\mathbb{R}^d$ under the additive log ratio transformation

$$y_i = \log(x_i/x_p);\ i = 1, 2, \ldots, d$$

Conversely, any vector which lies in $\mathbb{R}^d$ may be transformed to $\mathbb{S}^d$ by the additive logistic transformation

$$x_i = e^y_i/(1 + \sum_{j=1}^{d} e^{y_j}); i = 1, 2, \ldots, d$$

If $\mathbf{y}$ follows a multivariate normal distribution, $MVN(\boldsymbol{\mu}, \Sigma)$, on $\mathbb{R}^d$, then it follows that $\mathbf{x}$ follows a Logistic Normal distribution, $\mathcal{L}(\boldsymbol{\mu}\ , \Sigma)$, on $\mathbb{S}^d$. Thus the pdf of $\mathbf{x}$ may be expressed as:

$$\begin{aligned} f(\mathbf{x}) &= f(x_1, x_2, \ldots, x_{d+1}) = f(x_1, x_2, \ldots, x_d) = f(\mathbf{y})|\mathbf{J}_{\mathbf{x}\to\mathbf{y}}| = f(\mathbf{y})\frac{1}{\prod_{j=1}^{p} x_j} \\ &= (2\pi)^{-d/2}|\mathbf{\Sigma}\ |^{-1/2} \left(\prod_{i=1}^{p} x_i\right)^{-1} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}\ )^T \Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu}\ )\right\} \quad (\mathbf{x} \in \mathbb{S}^d, \mathbf{y} \in \mathbb{R}^d) \end{aligned}$$

Aitchison (2003) outlines the advantages of using this distribution to model compositional data, over the Dirichlet class of distributions. This distribution can model both dependent and independent structures in the data and also, test hypothesis on independence. Modeling the data as Logistic Normal provides access to the well-developed methods of statistical analysis for the normal distribution, such as estimation and hypothesis testing of the model parameters, tests of normality and multivariate analysis.

3.3.1 Fitting a Logistic Normal Model

We apply the additive log ratio transform to the compositional data, which transforms it to $\mathbb{R}^d$ space. We then fit a multivariate normal distribution to the transformed data.

The maximum likelihood parameter estimates for the multivariate normal distribution $MVN(\boldsymbol{\mu}, \Sigma)$, given observed data $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are given by

$$\hat{\boldsymbol{\mu}}\ = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$$

and

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\boldsymbol{\mu}})(\mathbf{y}_i - \hat{\boldsymbol{\mu}})^{\mathsf{T}}$$

We use the mclust library, designed by Fraley and Raftery (2006), within the R Development Core Team (2005) software package to calculate these maximum likelihood estimates for each group in the data. We make the assumption of homogeneity of covariance across all groups. The estimated parameters are then used as the maximum likelihood parameter estimates for the Logistic Normal distribution fitted to the compositional data.

3.4 Model Comparison

To compare the fit of the Dirichlet, Nested Dirichlet and Logistic Normal models, we use a variant of the BIC proposed by Steele (2002), which is an adjustment to that proposed by Schwarz (1978). The latter form of the BIC penalizes each parameter in the model by $\log n$, where n is the overall sample size. This form assumes that all observations contribute to each parameter estimate in the model. The adjusted BIC proposed by Steele (2002) is preferable for models in which some or all of the parameters are group dependent:

$$\begin{aligned}\text{BIC}_{\text{adj}} &= 2\text{LL} - k\log n - \sum_{g=1}^{G} k_g \log n_g \\ &= 2\text{LL} - k_{\text{tot}}\log n - \sum_{g=1}^{G} k_g \log \frac{n_g}{n} \\ &= \text{BIC} - \sum_{g=1}^{G} k_g \log \frac{n_g}{n}\end{aligned}$$

where n_g is the number of observations in group g, k is the number of parameters estimated using the whole data set, k_g is the number of parameters estimated solely from data from group g and k_{tot} is the total number of parameters in the model.

For the both the Dirichlet and the Nested Dirichlet models, all parameters are estimated independently in each group. For the Dirichlet, we estimate p parameters for each group and for the Nested Dirichlet we estimate $q = p + k$ for each group. Therefore, for these models, the adjusted BIC is:

$$\begin{aligned}\text{BIC}_{\text{Dir}} &= 2\text{LL} - \sum_{g=1}^{G} k_g \log n_g \\ &= 2\sum_{g=1}^{G}\text{LL}_g - \sum_{g=1}^{G} k_g \log n_g \\ &= \sum_{g=1}^{G}(2\text{LL}_g - k_g \log n_g) = \sum_{g=1}^{G}\text{BIC}_g\end{aligned}$$

which is simply the sum of the BICs for each group.

For the Logistic Normal model, the parameters to be estimated are:

Therefore, for the Logistic Normal model, we have $k = \frac{(p)(p-1)}{2}$ and $\sum_{g=1}^{G} k_g = G(p-1)$ and the BIC is:

Parameter	Number of Parameters	Group dependent
$\boldsymbol{\mu_g}$	$(p-1) = d$	Yes
$\boldsymbol{\Sigma}$	$\frac{(p)(p-1)}{2} = \frac{d(d+1)}{2}$	No

$$\mathrm{BIC}_{\mathrm{LN}} = 2\mathrm{LL} - \left(\frac{p^2-p}{2} + G(p-1)\right)\log n - (p-1)\sum_{g=1}^{G}\log\left(\frac{n_g}{n}\right)$$

Thus, for the Logistic Normal model, we adjust the BIC by subtracting $(p-1)\sum_{g=1}^{G}\log(n_g/n)$ from the ordinary BIC.

3.5 Classification of Samples

Suppose we fit model $\mathcal{M}$ to a dataset with G groups, estimating a parameter set θ_g for each group, we may develop a rule for classification of observation x_i using the posterior probabilities of the observation belonging to each group, derived from Bayes rule. This posterior is expressed as:

$$p(z_{ig} = 1|x_i) = \frac{\tau_g p(x_i|\theta_g)}{\sum_{j=1}^{G}\tau_j p(x_i|\theta_j)}$$

where

- $z_{ig} = \begin{cases} 1 & \text{if } x_i \text{ belongs to group } g \\ 0 & \text{otherwise} \end{cases}$
- $\tau_g = p(z_{ig} = 1)$ the prior probability of belonging to group g

To test this classification rule on our data, we use cross-validation of our samples to remove the bias which would result from using the same samples to both develop and test the classification rule. Thus, we remove observation x_i from the sample, estimate model parameters θ_g for each group g, then calculate the posterior probabilities of x_i belonging to each group. We repeat this process for each observation in the sample and use the resulting classifications to test the model's classification ability. In our models, we assume there is no prior belief about group memberships, so we let $\tau_g = 1/G$ for $g = 1, 2, \ldots, G$. Whilst it is possible to estimate τ_g, we do not do this here because the group proportions in the training data are not representative of the group proportions in the population of interest.

4 Results

We fit the Dirichlet, Nested Dirichlet and Logistic Normal models to the two glycan HILIC data sets described in Section 2. The model fitting procedures are outlined in Section 3. For each data set, we classify observations into the predefined groups using cross-validation for each model. We quantify the agreement of the classifications with the true group memberships using two measures - the proportion of observations on the diagonal of the classification contingency table, as well the Rand Index proposed by Rand (1971). We also measure the model goodness of fit using the adjusted BIC, as discussed in Section 3.4. The search algorithm used to fit the nesting trees for the Nested Dirichlet models is detailed in Section 3.2.1.

4.1 A Two Group Model for the Lung Cancer Data

In this section, we apply our compositional data methods to the lung cancer data described in Section 2.1. Due to the apparently poor differentiation between the stages of lung cancer, we merge all lung cancer stages into a single group, which results in the data set being divided into 100 cancer cases and 84 control cases. We fit the Nested Dirichlet model for the set of 17 compositional variables.

Table 3: Cross-validated classifcation results for (a) the Dirichlet (b) the Nested Dirichlet and (c) the Logistic Normal models fitted to the lung cancer data set

	Cancer	Control
Cancer	63	37
Control	21	63

(a)

	Cancer	Control
Cancer	72	28
Control	15	69

(b)

	Cancer	Control
Cancer	81	19
Control	8	76

(c)

The nesting tree for the Nested Dirichlet data is estimated using all 184 observations (as opposed to allowing the nesting tree to differ across groups). The estimated tree structure is shown in Figure 3. This tree structure is constrained to be the same across groups and we estimate Nested Dirichlet parameters within each group. We also fit the Dirichlet and Logistic Normal models to the data and compare the three models. The contingency tables for the classifications are displayed in Table 3, along with measures of performance for the classifications resulting from each model, in Table 4.

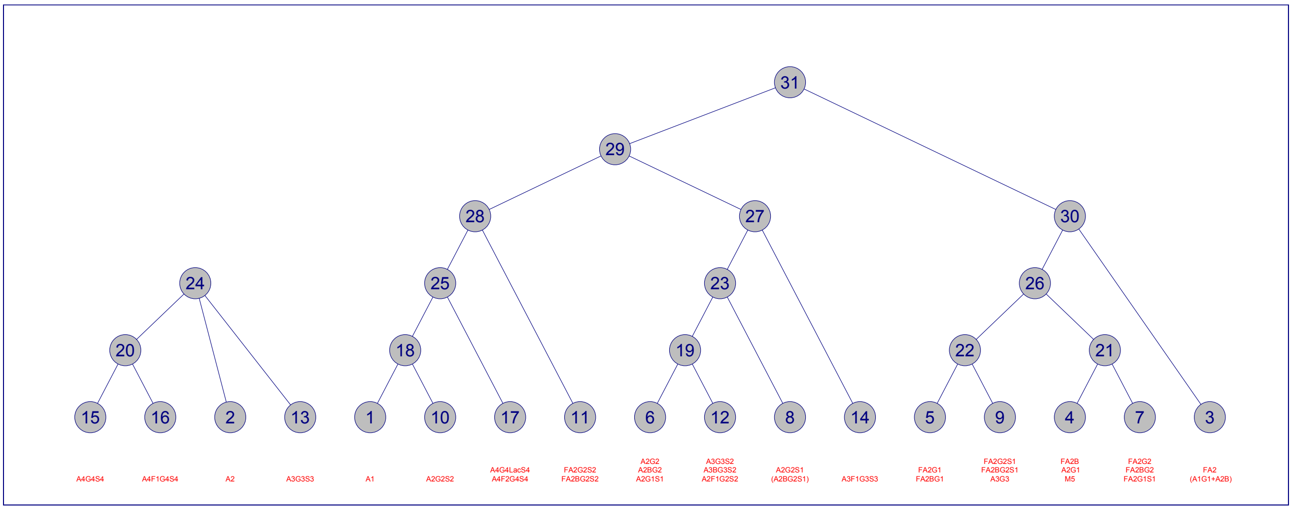

Figure 3: Nested Dirichlet nesting tree for the lung cancer data

The Nested Dirichlet model outperforms the Dirichlet model in terms of classification performance. However, the Logistic Normal model gives the best differentiation between control and cancer groups, of the three models. It has a misclassification rate of just under 15%.

Table 4: Classification diagnostics for the lung cancer data

	Dirichlet	Nested Dirichlet	Logistic Normal
Classification Rate	0. 68	0.77	0.85
Rand Index	0.57	0.64	0.75

The goodness of fit statistics are presented in Table 5. This table provides the adjusted BIC for each model. The BIC values are reflective of the classification results, as the Logistic Normal model has the largest BIC, whilst the Dirichlet has the smallest.

Table 5: Adjusted BIC for each model fitted to the Lung Cancer Data

	$\mathrm{BIC}_{\mathrm{adj}}$
Dirichlet	21783.77
Nested Dirichlet	23758.09
Logistic Normal	25155.66

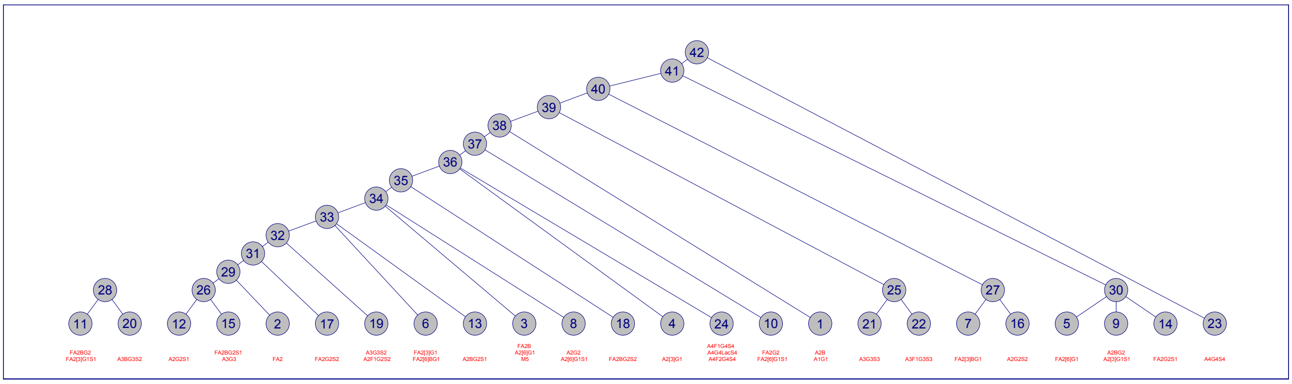

Figure 4: Nested Dirichlet nesting tree for the ovarian cancer data (control/cancer groups)

4.2 A Two Group Model for the Ovarian Cancer Data

We apply our compositional models to the ovarian cancer data outlined in Section 2.2. We combine cases in the Control and Benign groups and also cases with ovarian cancer (ovarian cancer + other gynecological cancers metastasizing to the ovaries). The models are applied to these two groups only. Our motivation for this is to test the ability of the models to differentiate between those with and without ovarian cancer. Group membership information is shown in Table 6.

Table 6: Data included in the control/cancer models for the ovarian cancer data set

Disease Status	**Cases**
Control/Benign (CB)	15
Ovarian Cancer/ Other gynecological cancers metastasizing to the ovaries (OC)	18

Table 7: Cross-validated classifcation results for (a) the Dirichlet (b) the Nested Dirichlet and (c) the Logistic Normal models fitted to the ovarian cancer data set (control/cancer groups)

	OC	CB
OC	12	6
CB	5	10

(a)

	OC	CB
OC	14	4
CB	4	11

(b)

	OC	CB
OC	11	7
CB	2	13

(c)

We fit the Nested Dirichlet distribution to the data, with the optimized nesting tree as shown in Figure 4. We also fit the Dirichlet and Logistic Normal models to the data. For each model, we classify observations using cross-validated classification methods. The agreement of these classifications with the true group memberships are detailed in Table 7. Measures of the classification performance may be found in Table 8.

For these data, the Nested Dirichlet is the best of the three models, in terms of classification performance. It has a misclassification rate of approximately 24%. It is closely followed by the Logistic Normal model, with a misclassification rate of just over 27%.

Table 9 shows the goodness-of-fit measures calculated for our fitted models. Here, the BIC for the Nested Dirichlet model is much larger than that for the Logistic Normal, which has a negative BIC.

Table 8: Classification diagnostics for the ovarian cancer data set (control/cancer groups)

	Dirichlet	Nested Dirichlet	Logistic Normal
Classification Rate	0.67	0.76	0.73
Rand Index	0.54	0.62	0.59

Table 9: Adjusted BIC for each model fitted to the Ovarian Cancer Data (control/cancer groups)

	BIC_{adj}
Dirichlet	81856.33
Nested Dirichlet	1486944.00
Logistic Normal	-1835.56

Thus, it appear that for these data, the Nested Dirichlet model is preferable to both the Dirichlet and the Logistic Normal, with regard to both goodness of fit and classification performance.

4.3 A Three Group Model for the Ovarian Cancer Data

We also define a three group data structure for the ovarian cancer data introduced in 2.2. We omit the borderline group and the group with other gynecological cancers metastasizing to the ovaries, as these groups are not mutually exclusive from other groups. We also omit the control and PPC groups, due to their small sample sizes. This divides our data into three groups, as outlined in Table 10 .

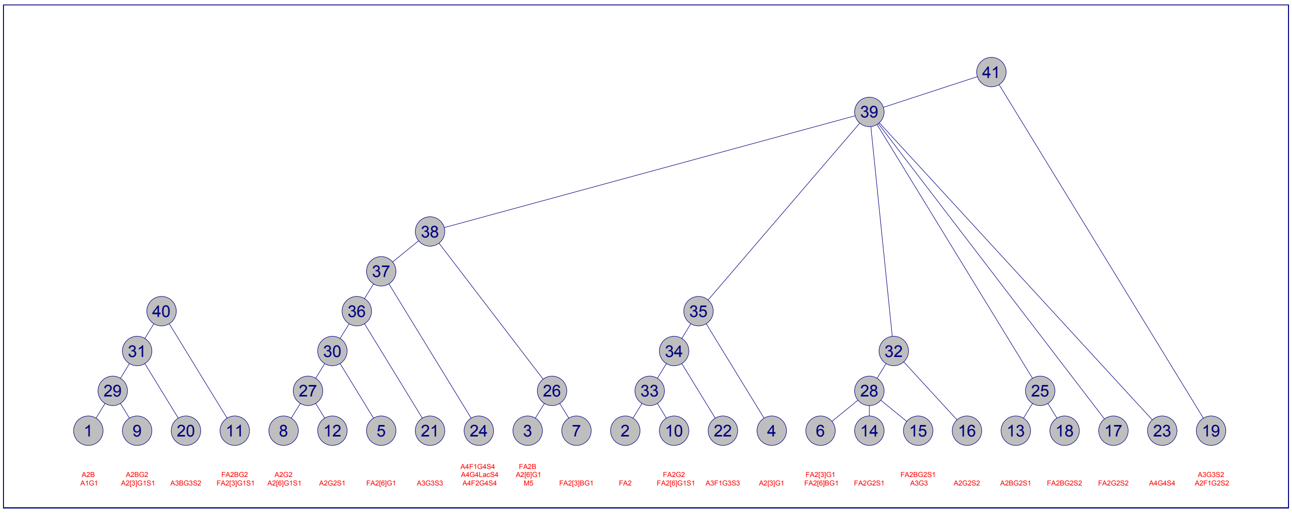

Figure 5: Nested Dirichlet nesting tree for the ovarian cancer data (3 groups)

We fit a Nested Dirichlet nesting tree to the three groups in this data set and the estimated tree structure is shown in Figure 5. This nesting tree has 17 nests, thus the fitted Nested Dirichlet distribution is the product of 17 Dirichlet distributions.

Table 10: Re-grouping of the Ovarian Cancer Data set into three groups

Disease Status	**Cases**
Control	OMITTED
Benign	12
Borderline	OMITTED
Ovarian Cancer	13
Primary peritoneal carcinomatosis (PPC)	OMITTED
Other gynecological cancers metastasizing to ovary (Other MTO)	OMITTED
Other gynecological cancers (Other)	21

In addition to the Nested Dirichlet model, we also fit the Dirichlet and Logistic Normal models to these data. For each model, we compute cross validated posterior group probabilities for each observation and classify the observation into the group with highest probability. The classification results are given in Table 11.

Table 11: Cross-validated classifcation results for (a) the Dirichlet (b) the Nested Dirichlet and (c) the Logistic Normal models fitted to the ovarian cancer data set (3 groups)

	Benign	Other	Ovarian Cancer
Benign	5	6	1
Other	8	11	2
Ovarian Cancer	4	5	4

(a)

	Benign	Other	Ovarian Cancer
Benign	5	6	1
Other	2	13	6
Ovarian Cancer	1	6	6

(b)

	Benign	Other	Ovarian Cancer
Benign	5	6	1
Other	6	10	5
Ovarian Cancer	1	6	6

(c)

None of the three models perform especially well, suggesting that there is perhaps not enough group information in the data. The Nested Dirichlet model performs better than the Dirichlet and the Logistic Normal models, in terms of both classification performance and goodness-of-fit. The assessment of classification performance for each model is given in Table 12.

Table 12: Classification diagnostics for the ovarian cancer data set (3 groups)

	Dirichlet	Nested Dirichlet	Logistic Normal
Classification Rate	0.43	0.52	0.46
Rand Index	0.54	0.55	0.55

Table 13 provides the BIC values for each model. The results are similar to those for the two group model. The Nested Dirichlet has the largest BIC and the Logistic Normal has the smallest.

Table 13: Adjusted BIC for each model fitted to the Ovarian Cancer Data (3 groups)

	BIC_{adj}
Dirichlet	139986.20
Nested Dirichlet	1527874.00
Logistic Normal	-1761.13

5 Discussion

This paper introduces the application of compositional data methods for the analysis of N-glycan HILIC data. The compositional models give good classification performance and their suitability can be compared using the BIC, which generally appears to select models with high classification rates. We hope to develop an R package (R Development Core Team, 2005) for use with glycan HILIC data, which can be used to fit these models and compare their performance.

5.1 Comparison of the Dirichlet model with the Nested Dirichlet model

The Nested Dirichlet distribution gives a much improved classification performance over the ordinary Dirichlet distribution. This is the case for all models fitted here. For those who wish to retain the convenience of modelling compositional data in the Simplex space, without losing the ease of model interpretion which this approach affords, the Nested Dirichlet distribution bears significant advantage over the Dirichlet distribution. It also appears to provide a better fit for the data, indicated by the larger BICs we observed for these models.

5.2 Comparison of the Nested Dirichlet with the Logistic Normal model

For modelling the Lung Cancer data set, the Logistic Normal model proves superior to the Nested Dirichlet model in terms of both classification performance and goodness-of-fit. However, the contrary is the case for the Ovarian Cancer data, for which the Nested Dirichlet model provides the best fit for the data. From this, we conclude that choice of model for compositional data should be made by examining the possibilities and selecting the model which seems to give the best results.

5.3 Computational Efficiency

Searching for the Nested Dirichlet tree structure proved to be cumbersome. Null (2008) suggests looking at all possible nesting trees and choosing that with the maximum likelihood. For p compositional variables, there are p^{p-2} arrangements of nesting tree possible. Since our data typically consists

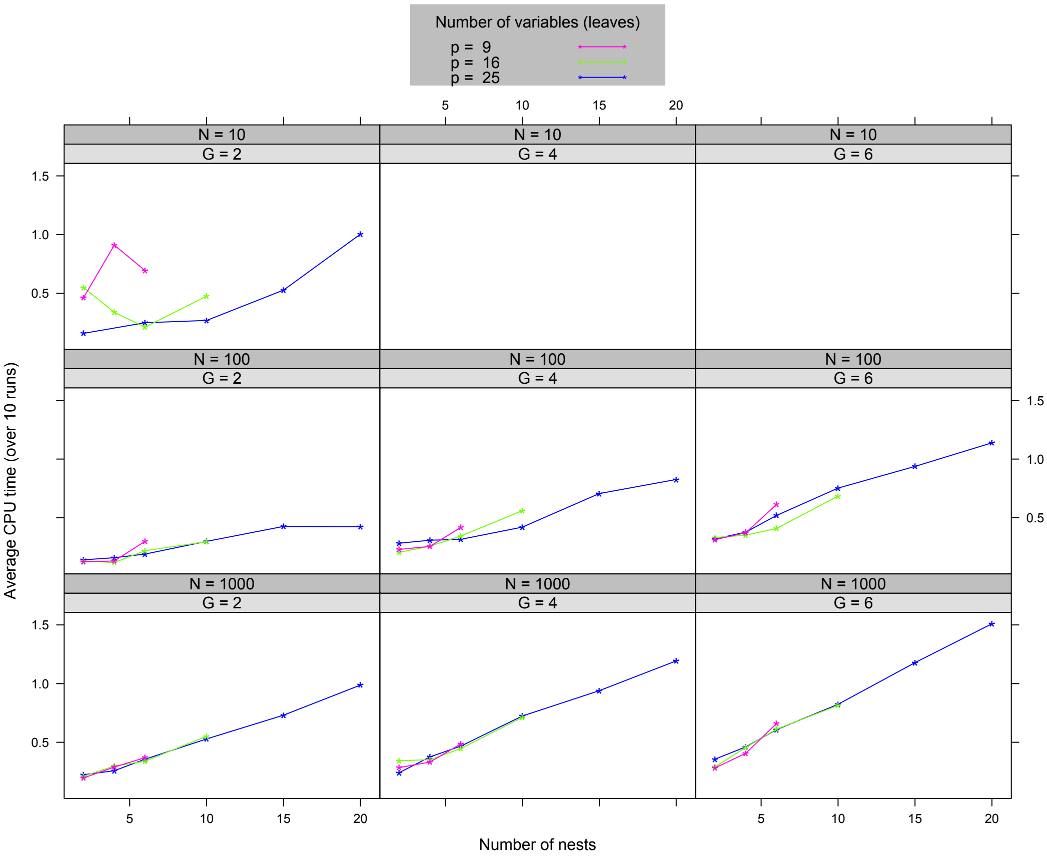

Figure 6: Recorded CPU time taken to fit Nested Dirichlet distributions, with varying parameters p, k, N and G

of $p = 17$ or $p = 24$, this is not computationally feasible. Therefore, we are required to employ some heuristic search strategies for identifying the correct tree. We tried three different search methods - a greedy search, simulated annealling and the forwards search method described in Section 3.2.1. The latter seemed to give the best result in general, that is the tree with largest BIC, so we adopt this approach for tree fitting. However, for the Nested Dirichlet distribution, the model fitting time increases approximately linearly with the number of nests in the tree, as would be expected (for every nest added, an additional Dirichlet model is required). This means that as the tree gets more complex, the search algorithm becomes very slow. Thus, with regard to efficiency, both the Dirichlet model and the Logistic Normal model are much faster to estimate than the Nested Dirichlet distribution.

References

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data* (Second ed.). The Blackburn Press.

Aitchison, J. and S. M. Shen (1980). Logistic-normal distributions: Some properties and uses. *Biometrika 67*(2), 261–272.

Arnold, J. N., , M. C. Galligan, T. Murphy, Y. Mimura-Kimura, J. Telford, A. Godwin, and P. M. Rudd (accepted 2011). Novel glycan biomarkers for the detection of lung cancer. *Journal of Proteome Research (In Press)*.

Arnold, J. N., R. Saldova, U. M. A. Hamid, and P. M. Rudd (2008). Evaluation of the serum N-linked glycome for the diagnosis of cancer and chronic inflammation. *Proteomics 8(16)*, 3284–93.

Bigge, J. C., T. P. Patel, J. A. Bruce, P. N. Goulding, S. M. Charles, and R. B. Parekh (1995). Nonselective and efficient fluorescent labeling of glycans using 2-amino benzamide and anthranilic acid. *Analytical Biochemistry 230*(2), 229–38.

Connor, R. J. and J. E. Mosimann (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association 64*(325), 194–206.

Ercan, A., J. Cui, D. E. W. Chatterton, K. D. Deane, M. M. Hazen, W. Brintnell, C. I. O'Donnell, L. A. Derber, M. E. Weinblatt, N. A. Shadick, D. A. Bell, E. Cairns, D. H. Solomon, V. M. Holers, P. M. Rudd, and D. M. Lee (2010, August). Aberrant IgG galactosylation precedes disease onset, correlates with disease activity, and is prevalent in autoantibodies in rheumatoid arthritis. *Arthritis & Rheumatism 62*(8), 2239–2248.

Fraley, C. and A. E. Raftery (2006). *MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering* (Technical Report 504 (revised December 2009) ed.). Department of Statistics: University of Washington.

Huang, J. (2005). Maximum likelihood estimation of Dirichlet distribution parameters. Technical report, Carnegie Melon University, Pittsburgh, Robotics Institute, School of Computer Science.

Kuster, B., S. F. Wheeler, A. P. Hunter, R. A. Dwek, and D. J. Harvey (1997). Sequencing of N-linked oligosaccharides directly from protein gels: in-gel deglycosylation followed by matrix-assisted laser desorption/ionization mass spectrometry and normal-phase high-performance liquid chromatography. *Analytical Biochemistry 250*(1), 82–101.

Kyselova, Z., Y. Mechref, P. Kang, J. A. Goetz, L. E. Dobrolecki, G. W. Sledge, L. Schnaper, R. J. Hickey, L. H. Malkas, and M. V. Novotny (2008). Breast cancer diagnosis and prognosis through quantitative measurements of serum glycan profiles. *Clinical Chemistry 54*(7), 1166–1175.

Minka, T. P. (2000;, February). Estimating a Dirichlet distribution. Technical report, M.I.T.

Null, B. (2008). The Nested Dirichlet distribution: Properties and applications. Working paper. Department of Management Science and Engineering, Stanford University.

Null, B. (2009). Modeling baseball player ability with a Nested Dirichlet distribution. *Journal of Quantitative Analysis in Sports 5*(2).

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

Ronning, G. (1989). Maximum likelihood estimation of Dirichlet distributions. *Journal of Statistical Computation and Simulation 32*(4), 215–221.

Royle, L., M. P. Campbell, C. M. Radcliffe, D. M. White, D. J. Harvey, J. L. Abrahams, Y. G. Kim, G. W. Henry, N. A. Shadick, M. E. Weinblatt, D. M. Lee, P. M. Rudd, and R. A. Dwek (2008). HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. *Analytical Biochemistry 376*(1), 1 – 12.

Saldova, R., L. Royle, C. M. Radcliffe, U. M. Abd Hamid, R. Evans, J. N. Arnold, R. E. Banks, R. Hutson, D. J. Harvey, R. Antrobus, S. M. Petrescu, R. A. Dwek, and P. M. Rudd (2007). Ovarian cancer is associated with changes in glycosylation in both acute-phase proteins and IgG. *Glycobiology 17*(12), 1344–1356.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Steele, R. J. (2002). *Practical importance sampling methods for finite mixture models and multiple imputation.* Ph. D. thesis, University of Washington.

Taylor, M. E. and K. Drickamer (2003). *Introduction to Glycobiology*. Oxford University Press.

A A Fixed Point Iteration for the Dirichlet distribution

Given a starting approximation for the parameters, $\mathbf{w}$, we use a fixed point iteration to numerically maximize the log likelihood of the Dirichlet distribution. If $\mathbf{X}$ is an $N \times p$ matrix of Dirichlet data and $\mathbf{x}_i$ is the ith row of $\mathbf{X}$, the log likelihood of the Dirichlet distribution is written

$$\begin{aligned}
l(\mathbf{w};\mathbf{X}) &= \log \prod_{i=1}^{N} f(\mathbf{x}_i|\omega) \\
&= \sum_{i=1}^{N} \log f(\mathbf{x}_i|\omega) \\
&= \sum_{i=1}^{N} \log \left(\frac{1}{B(\mathbf{w})} \prod_{l=1}^{p} x_{il}^{\omega_l - 1} \right) \\
&= \sum_{i=1}^{N} \log \Gamma(\omega^*) - \sum_{i=1}^{N} \sum_{l=1}^{p} \log \Gamma(\omega_l) + \sum_{i=1}^{N} \sum_{l=1}^{p} (\omega_l - 1) \log x_{il} \\
&= N \log \Gamma(\omega^*) - N \sum_{l=1}^{p} \log \Gamma(\omega_l) + \sum_{i=1}^{N} \sum_{l=1}^{p} (\omega_l - 1) \log x_{il}
\end{aligned}$$

A lower bound on the gamma function is

$$\Gamma(x) \geq \Gamma(\hat{x}) \exp\left\{(x - \hat{x})\Psi(\hat{x})\right\}$$

where $\psi(s)$ is the digamma function, the log derivative of $\Gamma(s)$. At iteration t, letting $x = \Gamma\ (\sum_{l=1}^{p} \omega_l^t)$ and $\hat{x} = \Gamma\ (\sum_{l=1}^{p} \omega_l^{t-1})$, allows us to set the lower bound

$$\Gamma(\sum_{l=1}^{p} \omega_l^t) \geq \Gamma(\sum_{l=1}^{p} \omega_l^{t-1}) \exp\left\{ \left(\sum_{l=1}^{p} \omega_l^t - \sum_{l=1}^{p} \omega_l^{t-1} \right) \Psi(\sum_{l=1}^{p} \omega_l^{t-1}) \right\} \tag{1}$$

We use this directly to place a lower bound on the log likelihood at iteration t, given $\mathbf{w}^{t-1}$, the parameter update at iteration $t-1$.

$$\begin{aligned}
l(\mathbf{w}^t;\mathbf{X}) \geq\ & N \log \left[\Gamma(\sum_{l=1}^{p} \omega_l^{t-1}) \exp\left\{ \left(\sum_{l=1}^{p} \omega_l^t - \sum_{l=1}^{p} \omega_l^{t-1} \right) \Psi(\sum_{l=1}^{p} \omega_l^{t-1}) \right\} \right] \\
& - N \sum_{l=1}^{p} \log \Gamma(\omega_l^{new}) + \sum_{i=1}^{N} \sum_{l=1}^{p} (\omega_l^{new} - 1) \log x_{il} \\
& = N \log \Gamma(\sum_{l=1}^{p} \omega_l^{t-1}) + N(\sum_{l=1}^{p} \omega_l^t)\Psi(\sum_{l=1}^{p} \omega_l^{t-1}) - N(\sum_{l=1}^{p} \omega_l^{t-1})\Psi(\sum_{l=1}^{p} \omega_l^{t-1}) \\
& - N \sum_{l=1}^{p} \log \Gamma(\omega_l^t) + \sum_{i=1}^{N} \sum_{l=1}^{p} (\omega_l^t - 1) \log x_{il} \\
& = l_{\text{lower}}(\mathbf{w}^t;\mathbf{X})
\end{aligned}$$

This lower bound can be maximized iteratively over $t = 1, 2, \ldots$, until $\mathbf{w}$ converges, by differentiating with respect to each ω_k^t; $k = 1, 2, \ldots, p$ at iteration t, and setting equal to zero.

Note that since $\omega^* = g(\omega_m)$ and $\frac{d\omega^*}{d\omega_m} = 1$,

$$\frac{\partial \log \Gamma(\omega^*)}{\partial \omega_m} = \frac{\partial \log \Gamma(\omega^*)}{\partial \omega^*}\frac{\partial \omega^*}{\partial \omega_m} = \psi(\omega^*).$$

and then the gradient of the lower bound on the log likelihood, with respect to ω_k^t, is written

$$\frac{\partial l_{\text{lower}}(\mathbf{w}^t;\mathbf{X})}{\partial \omega_k} = N\Psi(\textstyle\sum_{l=1}^p \omega_l^{t-1}) - N\Psi(\omega_k^t) + \sum_{i=1}^n \log x_{ik} = 0$$

(since $\mathbf{w}^{t-1}$ is constant with respect to $\mathbf{w}^t$). Rearranging the above in terms of the update at iteration t,

$$\omega_k^t = \Psi^{-1}\left(\Psi(\sum_{l=1}^p \omega_l^{t-1}) + \frac{1}{N}\sum_{i=1}^n \log x_{ik}\right)$$

This algorithm is continued until all parameters converge. (Our convergence criteria is that the maximum difference $\underset{k}{\operatorname{argmax}} |\omega_k^{t-1} - \omega_k^t| \leq 0.00001$).

The inversion of the digamma function is carried out using Newton's Method. Let $s = \Psi\ (\sum_{l=1}^p \omega_l^{t-1}) + \frac{1}{N}\sum_{i=1}^n \log x_{ik}$, then to solve $\omega_k^t - \Psi^{-1}(s) = 0$ for $\mathbf{w^t}$, iterate over

$$\omega_k^{new} = \omega_k^{old} - \frac{\Psi(\omega_k^{old}) - s}{\Psi'(\omega_k^{old})}$$

Initial estimates for the algorithm above are given by

$$\omega_k^{init} = \Psi^{-1}(s) \approx \begin{cases} \exp(s) + 0.5 & \text{if } s \geq -2.22; \\ -\frac{1}{s+\gamma} & \text{if } s < -2.22 \end{cases}$$

which comes from the approximation to the digamma function

$$\Psi(x) \approx \begin{cases} \log(x - \frac{1}{2}) & \text{if } x \geq 0.6; \\ -\frac{1}{x} - \frac{1}{\gamma} & \text{if } x < 0.6 \end{cases}$$

B Derivation of the Nested Dirichlet distribution

The probability density function for the Nested Dirichlet may be derived directly from the density function of the Dirichlet distribution. Before we derive this density function, some further notation for the Nested Dirichlet is required.

Some Notation for the Nested Dirichlet Distribution

- $x_l \searrow x_{p+j}$ denotes that x_l nested under the jth nesting variable.
- $I_j = \{i : x_i \searrow x_{p+j}; j = 1, 2, \ldots, k\}$ the indices of variables in nest j. I_0 is the set of indices for the unnested variables and X_0 denotes the set of unnested variables.
- $\chi\ _0$ is the set of unnested variables

- $\boldsymbol{\chi}_j = \{x_l : x_l \searrow x_{p+j}\}$ the set of variables nested under the jth nesting
- $\tilde{\boldsymbol{\chi}}_j = \mathscr{C}(\boldsymbol{\chi}_j)$ the variables nested under x_{p+j}, constrained to be compositional by the closure operator
- $\mathbf{W}_j = \{\omega_l : l \in I_j\}$ represents the set of Nested Dirichlet parameters for variables in the jth nest
- $\mathbf{W}^*_j = \sum_{l \in I_j} \omega_l$ the sum of parameters of variables in the jth nest
- $|\boldsymbol{\chi}_j|$ is the number of variables nested under the jth nesting variable.

Then for a Nested Dirichlet distribution with formed from a composition of p parts and with k nesting variables, we may express the probability density function of the Nested Dirichlet function as a product of the Dirichlet density functions contained within it.

Since $x_{p+1}, \dots, x_{p+k}$ are completely determined by $x_1, x_2, \dots, x_n$, we can write

$$f(x_1, x_2, \dots, x_{n+k}) = f(x_1, x_2, \dots, x_n) f(x_{n+1}, \dots, x_{n+k} | x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n)$$

Expressing this in terms of the sets of variables in each nest, gives us

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= f(\boldsymbol{\chi}_0, \boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \dots, \boldsymbol{\chi}_k) \\ &= f(\boldsymbol{\chi}_0) f(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \dots, \boldsymbol{\chi}_k | \boldsymbol{\chi}_0) \\ &= f(\boldsymbol{\chi}_0) f(\boldsymbol{\chi}_1 | \boldsymbol{\chi}_0) f(\boldsymbol{\chi}_2 | \boldsymbol{\chi}_0, \boldsymbol{\chi}_1) \cdots f(\boldsymbol{\chi}_k | \boldsymbol{\chi}_0, \boldsymbol{\chi}_, \dots, \boldsymbol{\chi}_{k-1}) \\ &= f(\boldsymbol{\chi}_0) f(\boldsymbol{\chi}_1 | x_{p+1}) f(\boldsymbol{\chi}_2 | x_{p+2}) \cdots f(\boldsymbol{\chi}_k | x_{p+k}) \end{aligned}$$

Clearly, the unnested variables are Dirichlet distributed

$$\boldsymbol{\chi}_0 \sim Dir(\mathbf{W}_0)$$

while the conditional distributions for each nest may be derived from the distribution of their compositional form.

$$\tilde{\boldsymbol{\chi}}_j \sim Dir(\mathbf{W}_j)$$

and since $\boldsymbol{\chi}_j = \tilde{\boldsymbol{\chi}}_j x_{p+j}$

$$\begin{aligned} f(\boldsymbol{\chi}_j | x_{p+j}) &= f(\tilde{\boldsymbol{\chi}}_j x_{p+j} | x_{p+j}) |\mathbf{J}_{\boldsymbol{\chi}_\mathbf{j} \to \tilde{\boldsymbol{\chi}}_\mathbf{j}}| \\ &= f(\tilde{\boldsymbol{\chi}}_j) \left(\frac{1}{x_{p+j}} \right)^{|\boldsymbol{\chi}_j| - 1} \end{aligned}$$

where $|\mathbf{J}_{\boldsymbol{\chi}_\mathbf{j} \to \tilde{\boldsymbol{\chi}}_\mathbf{j}}|$ is the determinant of the Jacobian matrix resulting from the change of variable rule, outlined in Appendix **??**.

For any $x_l \in \boldsymbol{\chi}_j$ and any $y_m \in \tilde{\boldsymbol{\chi}}_j$,

$$\frac{\partial y_m}{\partial x_l} = \frac{\partial (x_m / x_{p+j})}{\partial x_l} = \begin{cases} \frac{1}{x_{p+j}} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Then

$$|\mathbf{J}_{\boldsymbol{\chi}_\mathbf{j} \to \tilde{\boldsymbol{\chi}}_\mathbf{j}}| = \begin{vmatrix} \frac{1}{x_{p+j}} & 0 & \dots & 0 \\ 0 & \frac{1}{x_{p+j}} & \vdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{x_{p+j}} \end{vmatrix} = \left(\frac{1}{x_{p+j}} \right)^{|\boldsymbol{\chi}_j| - 1}$$

Note that the dimension of the Jacobian is one less than the size of $\boldsymbol{\chi}_j$, since a Dirichlet distribution with d variables is fully determined by any subset of size $d-1$, as the pth variable degenerate. The Nested Dirichlet probability density function may easily be derived from the above.

$$
\begin{aligned}
f(\mathbf{x};\mathbf{w}) &= f(\boldsymbol{\chi}_0)f(\tilde{\boldsymbol{\chi}}_1)f(\tilde{\boldsymbol{\chi}}_2)\cdots f(\tilde{\boldsymbol{\chi}}_k)\left(\frac{1}{x_{p+1}}\right)^{|\boldsymbol{\chi}_1|-1}\cdots\left(\frac{1}{x_{p+k}}\right)^{|\boldsymbol{\chi}_j|-1}\\
&= \left(\frac{1}{B(\mathbf{W}_0)}\prod_{l\in I_0}x_l^{\omega_l-1}\right)\left(\prod_{j=1}^{k}\frac{1}{B(\mathbf{W}_j)}\prod_{l\in I_j}\left(\frac{x_l}{x_{p+j}}\right)^{\omega_l-1}\left(\frac{1}{x_{p+j}}\right)^{|\boldsymbol{\chi}_j|-1}\right)\\
&= \left(\prod_{s=0}^{k}\frac{1}{B(\mathbf{W}_s)}\right)\left(\prod_{l\in I_0}x_l^{\omega_l-1}\right)\left(\prod_{j=1}^{k}\left(\prod_{l\in I_j}x_l^{\omega_l-1}\right)\left(\frac{1}{x_{p+j}}\right)^{\sum_{l\in I_j}(\omega_l-1)+|\boldsymbol{\chi}_j|-1}\left(\frac{1}{x_{p+j}}\right)^{|\boldsymbol{\chi}_j|-1}\right)\\
&= \left(\prod_{s=0}^{k}\frac{1}{B(\mathbf{W}_s)}\right)\left(\prod_{l=1}^{p+k}x_l^{\omega_l-1}\right)\left(\prod_{j=1}^{k}\left(\frac{1}{x_{p+j}}\right)^{\sum_{l\in I_j}\omega_l-|\boldsymbol{\chi}_j|+|\boldsymbol{\chi}_j|-1}\right)
\end{aligned}
$$

Thus, the density of the Nested Dirichlet distribution is derived from the product of the $k+1$ distributions of each nest and results in the pdf

$$
f(\mathbf{x};\boldsymbol{\omega}) = \frac{\prod_{l=1}^{p+k}x_l^{\omega_l-1}}{\prod_{s=0}^{k}B(\mathbf{W}_s)\prod_{j=1}^{k}x_{p+j}^{\mathbf{W}_j^*-1}}
$$

Each variable l in the Nested Dirichlet model has a corresponding parameter w_l, which is identical to its Dirichlet parameter, in the distribution of its respective nest.

C Glycan Analysis

This section contains some brief information on the glycan data used in our statistical analysis. For each data set, we have identified the glycans which are predominantly found under each peak in the chromatograms. Details of these glycans are given for both the lung cancer data (see table 14) and the ovarian cancer data set (see table 15. Information on the nomenclature used for the glycans in this table is described in Section C.1. Further infomation may be found in Royle et al. (2008) and also on the website for the Dublin-Oxford Glycobiology group. In Section C.2, we provide a short summary on the methods used to release and separate the glycan structures from the serum samples from which these data were gathered.

C.1 Structural symbols for the N-glycans and their linkages and abbreviations used.

Briefly, all N-glycans have two core N-Acetylglucosamines (GlcNAc) and a trimannosyl core; F at the start of the abbreviation indicates a core fucose linked α 1- 6 to the core GlcNAc; A[y]a represents the number of antenna (GlcNAc) on the trimannosyl core linked to the α 1-y mannose arm; B, bisecting GlcNAc linked β 1-4 to core mannose; Fb after Aa represents the number b of fucose linked α 1-3 to antenna GlcNAc; Gc represents the number c of galactose linked $\beta 1-4$ on antenna; S(z)d represents the number d of sialic acids linked α 2-z to the galactose.

C.2 N-Glycan analysis

Release and purification of N-glycans from human serum

N-glycans were released from serum using the high-through-put method described by Royle et al. (2008). Briefly, serum samples were reduced and alkylated in 96-well plates, and then they were immobilized in SDS-gel blocks and were washed. The N-linked glycans were released using peptide N-glycanase F (1000 U/mL; EC 3.5.1.52) as described previously (Bigge et al., 1995; Kuster et al., 1997).

Fluorescent labeling of the reducing terminus of N-glycans

Glycans were fluorescently labeled with 2-aminobenzamide (2AB) by reductive amination (Bigge et al., 1995) (LudgerTag 2-AB labeling kit LudgerLtd., Abingdon, UK).

HILIC

HILIC was performed using a TSK-Gel Amide-80 4.6 x 250 mm column (Anachem, Luton, Bedfordshire, UK) on a 2695 Alliance separation module (Waters,Milford,MA) equipped with a Waters temperature control module and a Waters 2475 fluorescence detector. Solvent A was 50 mM formic acid which was adjusted to pH 4.4 with ammonia solution. Solvent B was acetonitrile. The column temperature was set to $30\,^{\circ}\mathrm{C}$. Gradient conditions were as follows: 60 min method (lung cancer cohort, Arnold et al. 2008)- a linear gradient of 35 to 47% solvent A over 48 min at a flow rate of 0.8 mL/min, followed by 1min at 47 to 100% A and 4min at 100% A, returning to 35% A over 1 min and then finishing with 35% A for 6 min; 120 min method (ovarian cancer cohort, Saldova et al. 2007) - a linear gradient of 26-52% A, over 104 min. at a flow rate of 0.4 mL/min. followed by 1 min. at 52-100% A and 6 min. at 100% A at 1 ml/min., returning to 26% A over 1 min. and then finishing with 26% A for 8 min. (Royle et al., 2008). Samples were injected in 80% acetonitrile (lung cancer cohort) or 74% acetonitryle (ovarian cancer cohort) as described by Saldova et al. (2007). Fluorescence was measured at 420 nm with excitation at 330 nm. The system was calibrated with a dextran ladder, as described previously (Saldova et al., 2007).

Table 14: Predominant glycans found under each peak for the Lung Cancer data set. The glycans in brackets are not major, but were identified from a 2 hour HILIC.

Peak	Predominant Glycans	Peak	Predominant Glycans
1	A1	10	A2G2S2
2	A2	11	FA2G2S2, FA2BG2S2
3	FA2, (A1G1, A2B)	12	A3G3S2, A3BG3S2, A2F1G2S2
4	FA2B, A2G1, M5	13	A3G3S3
5	FA2G1, FA2BG1	14	A3F1G3S3
6	A2G2, A2BG2, A2G1S1	15	A4G4S4
7	FA2G2, FA2BG2, FA2G1S1	16	A4F1G4S4
8	A2G2S1, (A2BG2S1)	17	A4G4LacS4, A4F2G3S4
9	FA2G2S1, FA2BG2S1, A3G3		

Table 15: Predominant glycans found under each peak for the Ovarian Cancer data set.

Peak	Predominant Glycans	Peak	Predominant Glycans
1	A2B, A1G1	13	A2BG2S1
2	FA2	14	FA2G2S1
3	FA2B, A2[6]G1, M5	15	FA2BG2S1, A3G3
4	A2[3]G1	16	A2G2S2
5	FA2[6]G1	17	FA2G2S2
6	FA2[3]G1, FA2[6]BG1	18	FA2BG2S2
7	FA2[3]BG1	19	A3G2S2, A2F1G2S2
8	A2G2, A2[6]G1S1	20	A3BG3S2
9	A2BG2,A2[3]G1S1	21	A3G3S3
10	FA2G2,FA2[6]G1S1	22	A3F1G3S3
11	FA2BG2,FA2[3]G1S1	23	A4G4S4
12	A2G2S1	24	A4F1G4S4, A4G4LacS4, A4F2G4S4