Compositional Data in Biomedical Research

Dean Billheimer
Vanderbilt University
Vanderbilt–Ingram Cancer Center
Nashville, TN, USA
dean.billheimer@vanderbilt.edu

Abstract

Modern methods of compositional data analysis are not well known in biomedical research. Moreover, there appear to be few mathematical and statistical researchers working on compositional biomedical problems. Like the earth and environmental sciences, biomedicine has many problems in which the relevant scientific information is encoded in the *relative abundance* of key species or categories. I introduce three problems in cancer research in which analysis of compositions plays an important role. The problems involve 1) the classification of serum proteomic profiles for early detection of lung cancer, 2) inference of the relative amounts of different tissue types in a diagnostic tumor biopsy, and 3) the subcellular localization of the BRCA1 protein, and it's role in breast cancer patient prognosis. For each of these problems I outline a partial solution. However, none of these problems is "solved". I attempt to identify areas in which additional statistical development is needed with the hope of encouraging more compositional data analysts to become involved in biomedical research.

Key words: mass spectrometry, ELISA, BRCA1, proteomics, Markov chain Monte Carlo

1 Introduction

Methods of compositional data analysis are almost unknown in the fields of biostatistics and biomedical research. A search using Google Scholar (www.scholar.google.com) shows about 550 citations of Aitchison's (1986) "The Statistical Analysis of Compositional Data". Only 10 of these are in medical or biomedical journals. This lack-of-use seems surprising since many methods in biostatistics make use of the *relative sizes* of key parameters, such as odds ratios, relative hazards, risks, and rates. Further, many of the emerging measurement technologies in biomedicine require a "normalization" of the measurements before any meaningful biological information can be extracted. Although the specific normalization techniques depend upon the technology, many share the trait that the relevant information is contained in the relative abundance of key chemical species or categories. As compositional data analysts, we know that when relative abundance is important, compositions are lurking.

In his CoDaWork'03 lecture Professor Aitchison (2003) takes the view that, "the challenge of solving practical problems should motivate our theoretical research." In the hope of following Prof. Aitchison's directive, I describe a number of problems from the field of cancer research involving the analysis of compositions. These problems are both real and unsolved (or, at best, partially solved). My goal is to motivate areas of theoretical research that I think are important in biomedical research, and to foster interest among compositional data researchers in biomedicine.

The problems are summarized as follows: 1) the detection of lung cancer based on serum proteomic profiles, 2) the statistical "unmixing" (source apportionment) of tissue types from tumor biopsies to improve molecular profiling of tumors, and 3) understanding the factors affecting subcellular localization of the BRCA1 protein, and its role in breast cancer biology and prognosis. For each

of these problems, I will describe a short scientific background, how the data are collected, and an approach to a problem solution. The amount of detail provided is roughly proportional to the amount of progress I've made the problem. Each of these problems is an on-going research project at the Vanderbilt–Ingram Cancer Center. If you have any good ideas, feel invited to contact me at the address shown above.

2 Detection of Lung Cancer via Serum Proteomics

with Pierre Massion, MD and Shuo Chen, MS
Similar to benchmark problem 2.

Matrix-assisted laser desorption/ionization, time-of-flight (MALDI–TOF) mass spectrometry (MS) is a leading measurement technology in proteomics Karas and Hillenkamp, 1988). This technique allows direct measurement of the protein signature of tissue, blood, or other biological samples, and holds tremendous promise for disease screening, diagnosis and treatment. Mass spectrometry produces highly structured data with *functional* form: each sampled mass per charge (m/z) value is paired with an observed intensity. The resulting spectra are characterized by many local maxima (peaks) that correspond to peptides and proteins present in the sample.

One potential area of application of this technology is the early detection of cancers from easily sampled biological specimens. Petracoin and others (2002) have proposed using MALDI–TOF MS with serum to detect ovarian cancer. However, much controversy has accompanied the use of this technology because of inherent nuisance variation in measurement, poor experimental design, and failure to reproduce results in independent experiments (see Baggerly and others, 2004; Ransohoff, 2005; for example). Figure 1 shows a typical MALDI–TOF spectrum obtained from human serum. The spectrum has been corrected for non-zero baseline, and the intensities have been rescaled to constant sum.

Work by Yildiz and others (2005, unpublished) identifies a region of the spectrum between 11500 Da/z and 11750 Da/z that is useful in segregating patients with lung cancer from normal controls. Initial analysis found five signals that assisted in discriminating between cancer cases and matched normal controls. These signals were later putatively identified to be different truncation forms of serum amyloid A (SAA) precusor and SAA isoform2. The signals occur at specific m/z locations in the spectrum, and will be denoted as "component 1", "component 2", ..., "component 5", arranged in increasing m/z order. SAA is believed to be an indicator of inflammation within the body. However, neither the function of this protein, nor the implications of its multiple truncation forms are fully understood. Yildiz has speculated that both the total amount of SAA and the relative amounts of the different forms may be an indicator of the presence of cancer.

For this study, we focus on the relative amounts of the five truncation forms of SAA as a possible discriminator for lung cancer. We use the relative heights (intensities) of the mass spectrum at the five putative molecular weights as indirect measures of concentration. Although intensity is not directly interpretable as concentration, I hypothesize that the relative heights of these signals may be more reliable than the absolute, normalized intensities (which are routinely used). By considering only the relative heights in a short region of the spectrum, we hope to by-pass some of the problems associated with baseline correction, normalization, and peak detection.

Figure 2 shows sample spectra from two subjects. The dotted vertical lines indicate the theoretical m/z values for the five SAA forms. The magnitude difference is typical of normalized spectra in this m/z range.

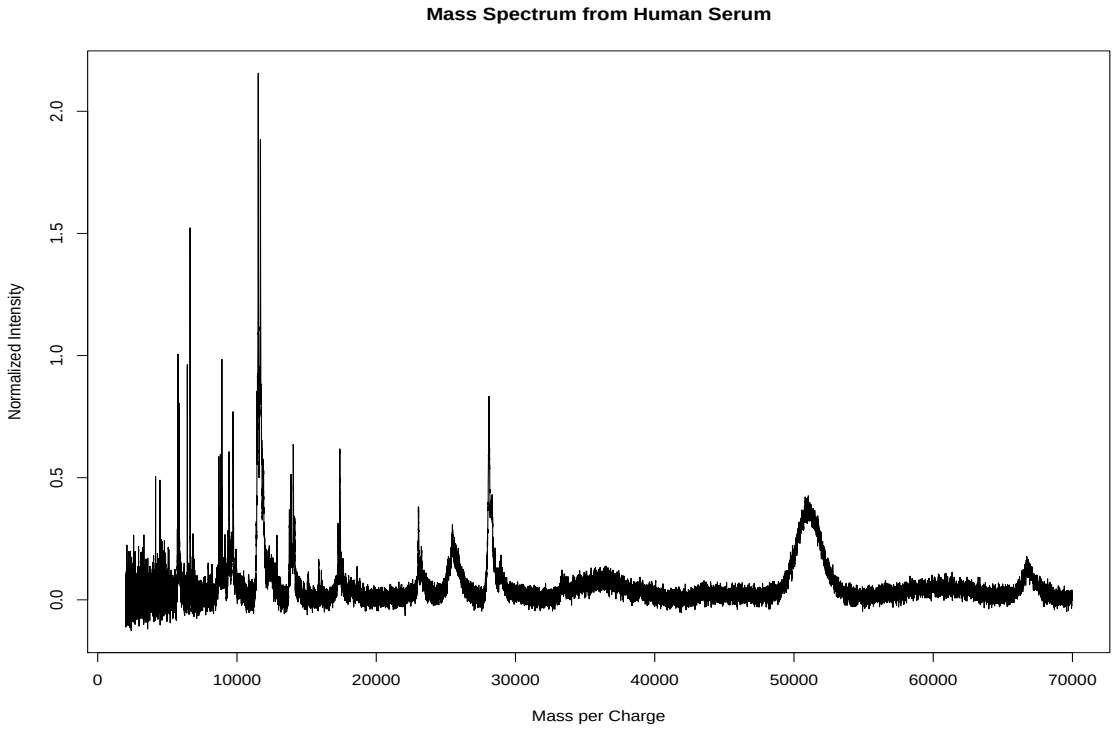

Figure 1: MALDI–TOF mass spectrum from human serum. Peaks in the spectrum correspond to proteins/peptides in the serum specimen.

2.1 Exploratory Data Analysis

Normalized intensities were acquired from the five signal m/z locations from each of 288 spectra (146 normal, 142 lung cancer). The intensities for each spectrum were then divided by their sum to form 5-part compositions. The data and group "means" (actually the location parameter under the logistic normal model. I will use this loose terminology throughout the paper.) are shown in Figure 3.

Although it is difficult to discern visually, the lung cancer patients may have slightly less of components 2 and 5, and slightly more component 4 than the normal patients. This is better displayed in Figure 4, showing only the estimated mean for each group.

The location parameter estimates for these groups are shown in Table 1. Despite the small mag-

Table 1: Normal and Lung Cancer composition location parameter estimates.

Group	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Normal	0.24	0.17	0.19	0.21	0.18
Lung Cancer	0.26	0.15	0.20	0.25	0.14

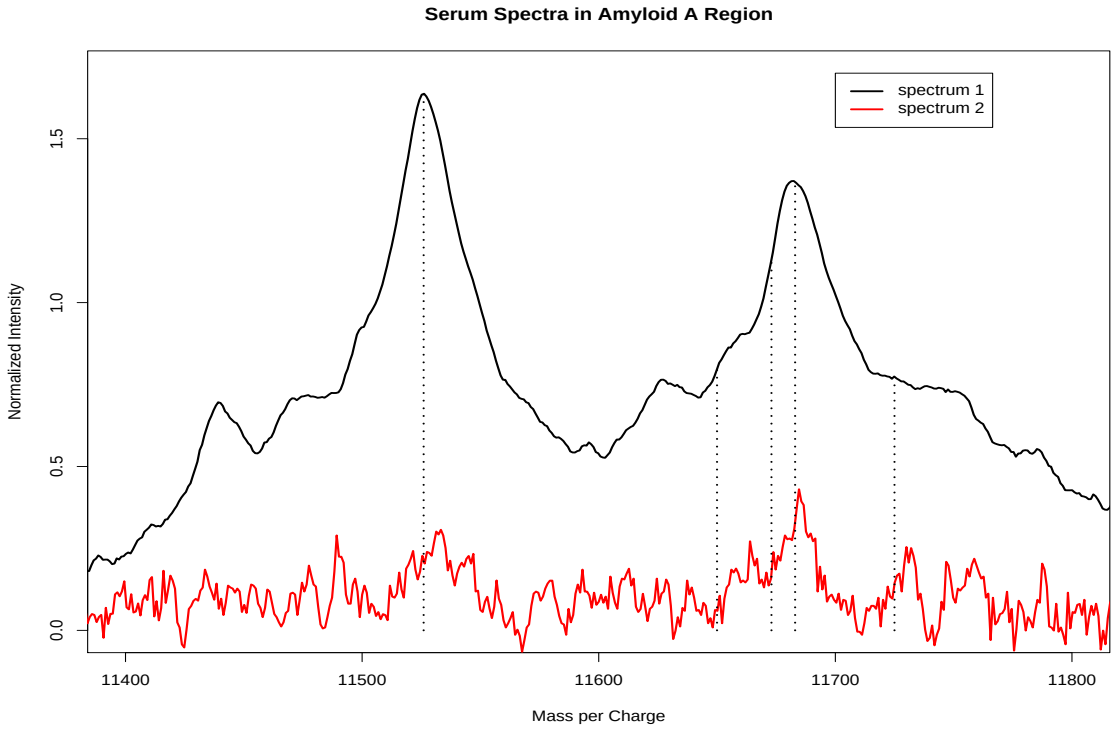

Figure 2: Mass per charge region containing the serum amyloid A truncation forms. Spectra from two subjects are shown to illustrate the variation. The dotted vertical lines indicate the theoretical m/z of the five truncation forms of SAA

nitude of differences in these proportions, MANOVA results for the alr transformed data indicate substantial differences in distribution for cancer versus normal comparison ($p < 0.0001$).

2.2 Classification of Compositions

The results of the last section suggest that we may be able to distinguish normal subjects from lung cancer patients based on their 5-part SAA compositions. We further explore this goal using a simple classifier based on the logistic normal distribution (Aitchison 1982)

To develop notation, let $\mathbf{z}$ denote a D-part composition

$$\mathbf{z} = (z_1, z_2, ..., z_D)', \quad \text{where} \quad z_i > 0, \quad \text{for all} \quad i = 1, 2, ..., D$$

and

$$\sum_{i=1}^{D} z_i = 1$$

Further, following Aitchison, define the additive logratio transformation (alr; $\phi(\cdot)$) as $\phi : \nabla^d \rightarrow \Re^{D-1}$

$$\phi(\mathbf{z}) = \left[\log\left(\frac{z_1}{z_D}\right), \log\left(\frac{z_2}{z_D}\right), \cdots, \log\left(\frac{z_{D-1}}{z_D}\right)\right]'$$

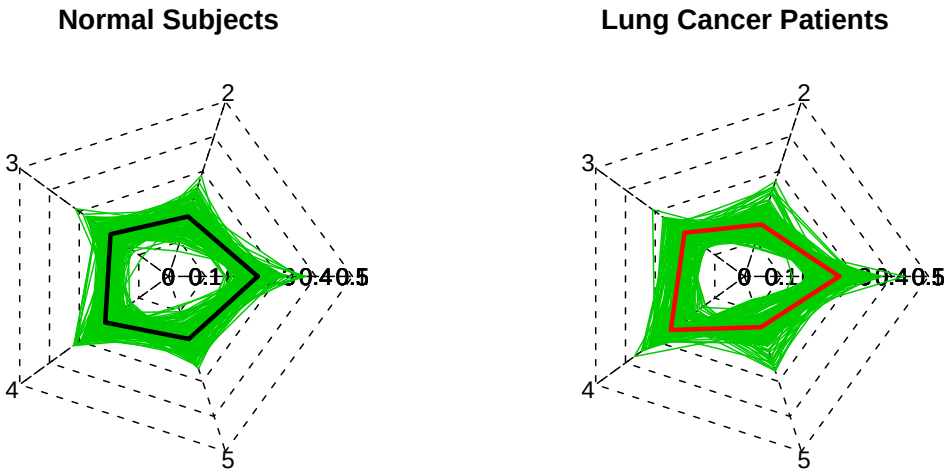

Figure 3: Five-part compositions for SAA truncation forms. The webplots show the individual spectrum compositions (green polygons), as well as the location parameter estimates for each group (black polygon for normal, red for cancer). The radials, numbered 1 to 5, denote the axes for each of the 5 SAA components.

Then the logistic normal density is given as follows:

$$f\left(\mathbf{z} \mid \boldsymbol{\mu}, \Sigma\right) = \left(\frac{1}{2\pi}\right)^{\frac{D-1}{2}} \mid \Sigma \mid^{-\frac{1}{2}} \left(\frac{1}{\prod_{i=1}^{D} z_i}\right) \exp\left[-\frac{1}{2}(\phi(\mathbf{z}) - \boldsymbol{\mu})^{'} \Sigma^{-1} (\phi(\mathbf{z}) - \boldsymbol{\mu})\right] \tag{1}$$

We abbreviate this using the following notation $\mathbf{z} \sim L^{D-1}\left(\boldsymbol{\mu}, \Sigma\right)$

To begin, we use a simple classifier based on Bayes rule

$$\frac{Pr(cancer \mid \mathbf{y})}{Pr(normal \mid \mathbf{y})} = \frac{f(\mathbf{y} \mid cancer)}{f(\mathbf{y} \mid normal)} \times \frac{Pr(cancer)}{Pr(normal)}$$

Thus for a single (multivariate) datum ($\mathbf{y}$), the posterior odds of cancer are equal to the likelihood ratio times the prior odds. For now, assume that the prior odds are one, and estimate the location and dispersion parameters (separately for cancer and normal groups) using the logistic normal distribution and maximum likelihood. This done, it is straight-forward to compute the odds (or probability) of belonging to the lung cancer group for each serum composition. We produce a quick summary by classifying each composition as "cancer" if its posterior probability is greater than 0.5. These results are shown in Table 2.

The overall classification accuracy is 67%. The specificity is not bad at 77%, but the sensitivity is not very good (56%). As a test of out-of-sample performance, we apply the classification model

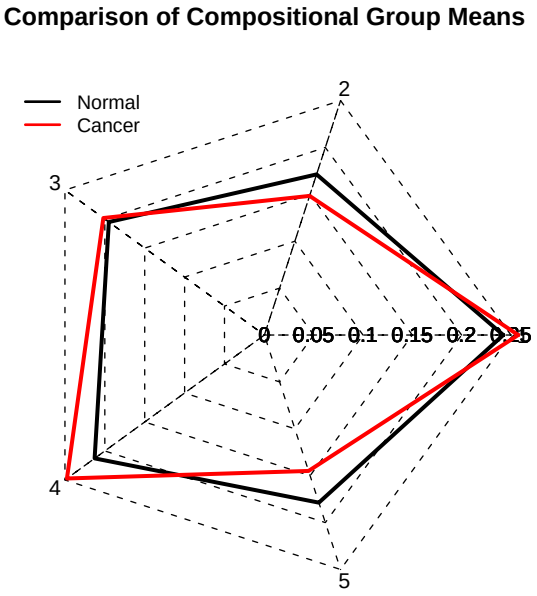

Figure 4: Mean compositions for normal and cancer groups (5-part SAA composition). The cancer group appears to have a slightly greater relative amount of component 4, and slightly less components 2 and 5. Normal mean $(.24, .17, .19, .21, .18)$, cancer mean $(.26, .15, .20, .25, .14)$

Table 2: Classification of Serum Compositions

Classification	Normal	Lung Cancer
Class Normal	63	113
Class Lung Cancer	79	33

fit above to an independent testing set of 137 serum specimens. These sera were collected at a different institution, and their spectra acquired at a different time than the original 288. The results of this testing are summarized in Table 3. The testing specificity is comparable to that seen in the training data set (70%). However, the testing sensitivity is only 39%, even worse than that in the training set. Figure 5 shows a "post-mortem" graph of the normal and cancer compositional means for both training and testing data sets.

The mean composition for normals in the test set shows good agreement with the corresponding compostion in the training set. However, the cancer group mean from the test set is similar to the normal means. The distances between group means is summarized in Table 4. This distance is computed via a norm under the Aitchison geometry in the simplex (Aitchison, 1992; Pawlowsky–Glahn and Egozcue, 2001; Billheimer, *et al.* 2001).

Table 3: Test Classification of Serum Compositions

Classification	Normal	Lung Cancer
Class Normal	23	63
Class Lung Cancer	10	41

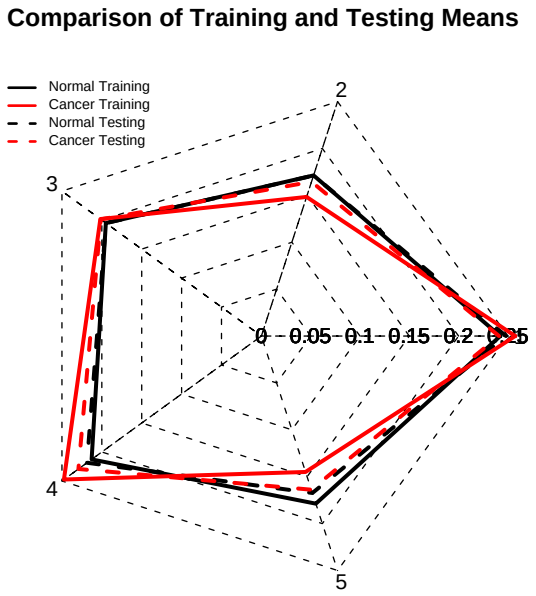

Figure 5: Mean compositions for normal and cancer groups (5-part SAA composition) for training and testing sets. The cancer mean in the testing set appears very similar to the normal mean compositions.

2.3 Summary and Open Questions

The post-mortem results indicate a problem with reproducibility of the lung cancer patient compositions in training and test data sets. The source of this discrepancy is unclear. One possibility is that the original training data were overfit, and that the five SAA signals are not a consistent marker of cancer–normal differences. A second possibility is that serum specimens from the testing cohort were handled differently than those in the training cohort. A systematic difference in specimen processing could account for the observed differences. Finally, it is possible that treating the five SAA intensities as compositions does not improve upon use of the measured intensities. Indeed, the SAA signals were identified from analysis based on normalized intensity.

Questions

You may notice several similarities with benchmark problem 2.

1. Are there better classifiers for compositions than one proposed? (based on Bayes rule with a logistic normal density).

Table 4: Distance Between Compositional Group Means

Group	Cancer Train	Normal Test	Cancer Test
Normal Train	0.300	0.075	0.125
Cancer Train		0.240	0.186
Normal Test			0.079

2. How should one choose signals to include/exclude in the composition? (There are at least 180 consistently expressed signal peaks from which to choose in the serum spectra.)

3. Is this kind of response/model selection different that in the unconstrained case?

4. How can one assess whether a compositional model is more/less appropriate than an intensity based model (i.e., one based on the non-constrained, measured intensities?)

3 "Unmixing" Tissue Protein Signatures from Tumor Biopsies

with Marta Guix, MD and Ray Mernaugh, PhD

For many types of cancer, it is common practice to acquire tissue biopsy to assist in disease diagnosis. A part of this tissue is examined by a pathologist to aid in the diagnosis. In several ongoing research studies at Vanderbilt Medical Center, a second part of the biopsy is used to assess the amounts of specific proteins that are believed important in cell growth/death, sensitivity to therapeutic drugs, or cancer metastasis. Because the (research) biopsy sample contains unknown amounts of different tissues types, the measured protein concentrations should be normalized to the relative amount of the tissue of interest.

For example, a breast tumor biopsy may contain normal, dysplastic and cancerous epithelial cells, as well as stromal components (fatty and connective tissue) and blood and lymphatic vessels. The percentage of each of this components changes as a function of many variables. Some of these are related to patient characteristics. These include age (for example: as a woman ages, the fibrous connective tissue is replaced by fatty tissue), hormonal status (pre-menopausal versus post-menopausal), and the different types of breast tumors themselves. Another major factor affecting the tissue composition in the biopsy is the physical acquisition of biopsy material. Some biopsies are more successful in hitting the center of the tumor target than others.

Each of the tissue components in a biopsy specimen produces a different protein "signature". In traditional immunohistochemistry (IHC) techniques, certain proteins have been used as staining targets. For example, citokeratins are characteristic of epithelial tissue, while collagen and fibronectin are found in the stroma. Once stained, the different tissue components can be visually identified by a pathologist. Typically, a pathologist focuses only on the cell type of interest (e.g., cancerous epithelium), and qualitatively scores the staining intensity (low, medium, high).

ELISA assays (described below) offer a more sensitive and quantatitive method to measure the protein concentrations from tissue. However, the measured concentrations will be based on the total protein contributed from all tissues in the mixture. To accurately assess the protein concentrations in the tissue types of interest, a method is needed to mathematically adjust for the relative amounts of the different tissue types in the biopsy. However, the amount of each tissue type is not directly observable. This is an example of the source apportionment problem that has been identified in many areas of science (see e.g., Rowe 2003).

Using the idea of a chemical mass balance, we consider the total tissue protein concentrations to be a weighted sum of the individual tissue signatures, where weights are determined by the relative amounts of the different tissue types. Our goal is to infer the protein expression in cancer epithelium, and relate measured levels of key proteins (e.g., ER, EGFR, p53, etc.) to clinical outcomes. Because the amounts of different tissue types, and the protein signatures of each are not known precisely, we are faced with a difficult inference problem.

3.1 ELISA Protein Assays

We measure the amount of individual proteins using enzyme-linked immunosorbent assays (ELISA). This is an immunology based measurement method to detect the presence of an antigen (e.g., a protein target) in a sample. It requires two antibodies (recognition units), each specific to different recognition sites on the antigen. One antibody is coupled to an enzyme that produces a fluorescent signal when bound to the target. The steps in a "sandwich" ELISA (www.wikipedia.org) are as follows:

1. Prepare a surface (plate) to which a known quantity of "capture" antibody is bound.
2. The antigen-containing sample is added and allowed to bind to the fixed antibody.
3. Wash the plate, so that unbound antigen is removed.
4. Apply the enzyme-linked "detection" antibodies which are also specific to the antigen.
5. Wash the plate, so that unbound enzyme-linked antibodies are removed.
6. Apply a chemical which is converted by the enzyme into a fluorescent signal.
7. View the result: if it fluoresces, then the sample contained antigen.

To estimate the amount of protein, the optical density or fluorescent intensity of the sample is interpolated against a standard curve (typically derived from a serial dilution of the target). A separate ELISA assay is performed for each protein of interest.

3.2 A Statistical Model

Billheimer (2001) proposed a statistical model based on a chemical mass balance and the logistic normal distribution (see Equation 1 above). The mass balance model may be written as

$$\mathbb{E}\left[\mathbf{Y}_i\right] = \sum_{j=1}^{p} \alpha_{ji}\,\boldsymbol{\theta}_j \;=\; [\boldsymbol{\theta}_1|\boldsymbol{\theta}_2|\ldots|\boldsymbol{\theta}_p] \begin{bmatrix} \alpha_{1i} \\ \alpha_{2i} \\ \vdots \\ \alpha_{pi} \end{bmatrix} \;=\; \boldsymbol{\Theta}\,\boldsymbol{\alpha}_i$$

and the logistic normal included as a perturbation error.

$$\mathbf{Y}_i = \boldsymbol{\theta}\,\boldsymbol{\alpha}_i \oplus \boldsymbol{\epsilon}_i \quad \text{for } \; i = 1, 2, ..., n \tag{2}$$

where,

$\mathbf{Y}_i$ is a $D \times 1$ vector of (relative) concentrations of the D protein species. $(i = 1, 2, ..., n)$

$\boldsymbol{\theta}_j$ is a $D \times 1$ vector of the protein profile for source j (p sources).

$\boldsymbol{\alpha}_i$ is a $p \times 1$ vector of mixing coefficients

$\boldsymbol{\epsilon}_i$ is unexplained variation (later modeled as $L^{D-1}(\boldsymbol{\mu}, \Sigma)$

Unlike the typical linear regression setting, $\boldsymbol{\Theta}$ and $\boldsymbol{\alpha}_i$ are unknown. As a consequence, the model is not identifiable. In addition, the statistical model presents a number of difficulties, including

- Physical mixing (convex combination) is not often addressed in the field of statistics (unlike mixture distributions).
- The convex combination combines compositions to produce another composition. However, this operation is (apparently) not included in the Aitchison simplicial geometry.
- If each $\mathbf{Y}_i$ is associated with a unique $\boldsymbol{\alpha}_i$, we have an instance of the incidental parameter problem (Neyman and Scott, 1948). This is problematic for large sample properties of maximum likelihood estimators.
- The mixing coefficients and tissue source profiles must satisfy positivity and summation constraints.
- The number of tissue types (sources) is not known.

We take the following approach toward forming a solution to this problem. Note that this approach relies on substantial knowledge of breast tissue biology, antibody use for molecular recognition, as well as statistics. We rely on each of the following assertions/constructions.

- The major tissue types likely to be present in a biopsy are known, and information (data) is available about the protein source profiles for each (pure) tissue type. This will allow the construction and use of informative prior distributions for the protein sources profiles.
- The mixing proportions (proportion of each tissue type) can be modeled as a draw from an unknown distribution. The distribution will be "informed" in the sense that we will use expert knowledge to provide approximate proportions of each tissue type anticipated in a biopsy. Note that this distribution can also be updated via examination of independent tissue biopsies.
- We may choose which proteins (antigens) to measure to achieve a balance between model identification and measurement of biologically interesting proteins. Some of the assays will be used to estimate the amount of "nuisance" tissue types, and some will be used to measure the proteins of medical/scientific interest. This is a trade-off between more precise inference and the number of relevant proteins to be measured.

We may combine these (informative) prior distributions with the logistic normal likelihood for inference in a Bayesian setting. A more complete model specification begins with Equation 2 above, and places logistic normal prior distributions on the $\boldsymbol{\theta}_j$ vectors (the "pure" tissue protein profiles), as well as the mixing parameter vector $\boldsymbol{\alpha}_i$ for $(i \in 1, 2, ...n)$. Billheimer (2001) constructs a similar model, and approximates the joint posterior distribution using Markov chain Monte Carlo (MCMC) sampling.

We are in the early stages of this project, and do not yet have data that we can share. However, we do have questions.

3.3 Questions

1. Can we identify "missing" tissue types? That is, tissues that are present in the sample, but missing from the model. (How does one assess model adequacy?) Will model "residuals" (differences between predicitive distributions and observations) provide information about the missing components?

2. How should we choose the which proteins to measure? We can only perform a fixed number of ELISA assays on each tumor specimen (probably about 5 to 8). We have competing priorities in normalization/model identification and scientifically interesting proteins.

4 Subcellular Localization of BRCA1 Protein

with Fen Xia, MD PhD

Women who carry BRCA1 mutations have an 85% risk of developing breast cancer by age 70, which about 20-fold higher than the general population. Accumulating evidence suggests that BRCA1 is required for many cellular processes including DNA replication and repair, cell-cycle checkpoints, transcription regulation, protein ubiquitination, and apoptosis (Feng, and others 2004). However, the mechanisms through which BRCA1 acts as a tumor supressor are less well-known. The elimination of one or more of these BRCA1 functions may lead to tumor development.

Recent evidence has shown that BRCA1 is a shuttle protein, which is actively transported between the cell nucleus and cytoplasm. Further, its localization within the cell has important consequences for BRCA1 function. Experiments in tissue culture indicate that wild-type P53 is required for BRCA1 nuclear export (Feng, and others 2004).

In this study we examine the expression and subcellular localization of BRCA1 in human breast cancer tumor tissue. In addition we seek to understand factors affecting localization of the BRCA1 protein, and its role in breast cancer biology and prognosis. To this end, 31 breast cancer tumors were prepared by standard IHC techniques, and stained for BRCA1 and a mutation of the P53 protein. P53 is a central protein in the apoptotic pathway (programmed cell–death pathway). A mutated form of P53 suggests loss of this cell regulation mechanism. For each tumor specimen, several hundred cancer epitheilial cells were painstakingly evaluated for the presence of P53 mutation, and localization of the BRCA1 protein. In each cell BRCA1 was categorized as in either 1) nucleus only, 2) nucleus and cytoplasm, 3) cytoplasm only, or 4) absent.

Our goals in this experiment are to 1) understand the relationship between P53 status and BRCA1 localization, and 2) evaluate whether BRCA1 localization is related to clinical outcomes (e.g. progression free surivial).

4.1 Statsitsical Modeling

The data are cell counts from 31 breast cancer tumors. We define cell groups based on P53 status: P53 negative denotes wild-type (WT) protein, while P53 positive denotes mutation (MUT). Within each P53 group, we consider the intracellular location of BRCA1 protein. These categories include nucleus only, nucleus and cytoplasm, cytoplasm only, and non-staining cells (non-detection of BRCA1). For most tumors 500 cancer cells were evaluated. A few tumor specimens (4) contained limited numbers of cancerous epitheilial cells. These specimens allowed evaluation of 150–300 cells per tumor.

We model cell counts in each of the four categories as conditionally multinomial, given an unobserved compositional vector parameter. The compositional "mean" parameter is in turn modeled using Aitchison's (1982) logistic normal density (Equation 1) in a hierarchical setting. Because each tumor contains both P53 positive and negative cells, we anticipate that paired (within tumor) analysis design may provide sharper inference by explicitly accounting for tumor-to-tumor variation.

Hence, write the cell count vector from P53 group j from tumor i, $\mathbf{y}_{ij}$, is conditionally multinomial given the latent composition vector $\boldsymbol{\theta}_{ij}$ ($\boldsymbol{\theta}_{ij} \in \nabla^d$ for i in $\{1, 2\}$ groups, and i in $\{1, 2, ..., 31\}$ tumors.

The $\boldsymbol{\theta}_{ij}$ compositions are modeled as independent draws from $L^{D-1}(\boldsymbol{\mu}_{ij}, \Sigma)$ where

$$\boldsymbol{\mu}_{ij} = \boldsymbol{\gamma}_i + \boldsymbol{\alpha}_j$$

Thus, $\boldsymbol{\gamma}_i$ accounts for tumor-to-tumor variation, while the $\boldsymbol{\alpha}_i$ describes group differences. Note that while $\boldsymbol{\mu}_{ij}$, $\boldsymbol{\gamma}_i$, and $\boldsymbol{\alpha}_j$ are defined in $\Re^{D-1}$, they are also interpretable as compositions via the inverse alr transform (see e.g., Billheimer, and others 2001). We complete model specification by assigning proper, but relatively uninformative conjugate prior distributions to Σ, $\boldsymbol{\gamma}_i$, and $\boldsymbol{\alpha}_j$. To summarize

$$\begin{aligned}
\mathbf{y}_{ij} \mid \boldsymbol{\theta}_{ij} &\sim \mathrm{Mn}(n_{ij}, \boldsymbol{\theta}_{ij}) \\
\boldsymbol{\theta}_{ij} \mid \boldsymbol{\gamma}_i, \boldsymbol{\alpha}_j, \Sigma &\sim L^{D-1}(\boldsymbol{\mu}_{ij}, \Sigma) \\
\boldsymbol{\gamma}_i &\sim \mathrm{N}_{D-1}(\boldsymbol{\mu}_\gamma, \Sigma_\gamma) \\
\boldsymbol{\alpha}_j &\sim \mathrm{N}_{D-1}(\boldsymbol{\mu}_\alpha, \Sigma_\alpha) \\
\Sigma &\sim \mathrm{Wishart}(\Omega, \mathrm{df})
\end{aligned}$$

In this setting the joint posterior distribution can be sampled via MCMC using Metropolis–Hastings steps to update $\boldsymbol{\theta}_{ij}$, and Gibbs steps for other parameters.

Typical choices for $\boldsymbol{\mu}_\alpha$ and $\boldsymbol{\mu}_\theta$ are

$$\boldsymbol{\mu}_\alpha = \boldsymbol{\mu}_\theta = 0_{D-1} \text{ and } \Omega = a\mathcal{N}$$

where

$$\mathcal{N} = I_{D-1} + \mathbf{j}_{D-1}\mathbf{j}_{D-1}^{'} \ ,$$

0_{D-1} is a $(D-1)$-vector of 0's, I_{D-1} is a $(D-1)$ identity matrix, and $\mathbf{j}_{D-1}$ is a $(D-1)$-vector of ones. This is equivalent to specifying a LN prior distribution for $\boldsymbol{\xi}_\cdot = \phi^{-1}(\boldsymbol{\mu}_\cdot)$, centered at $\mathcal{I}_{D-1}$ (for $D = 4$, (1/4, 1/4, 1/4, 1/4)). Setting the hyperparameter $a = 0.5$ allows the 95% prior probability contour for $\boldsymbol{\xi}_\cdot$ to reach at least 0.05 for each component.

4.2 Results

We fit the model described above using MCMC to sample from the joint posterios distribution. Preliminary MCMC runs indicated that the Markov chain mixed quite rapidly, and that convergence did not depend strongly on starting value. We deemed a burn-in of 100 MCMC iterations to be adequate, and collected realizations for an additional 2000 iterations. Post-run graphical diagnostics indicated that this run length was adequate. Checks via parallel Markov chains with different starting values confirmed initial results.

Table 5 shows the posterior mean compositions for BRCA1 localization with P53 WT and P53 MUT. These point estimates are shown graphically, along with a 95% credible interval (for P53

Table 5: Compositional Group Means

Group	Nucleus	Nuc/Cyto	Cytoplasm	BRCA1 Neg
P53 WT	0.13	0.14	0.06	0.66
P53 MUT	0.36	0.14	0.04	0.46

WT) in Figure 6. The black polygon, representing the P53 WT posterior mean indicates a very different composition from the P53 MUT posterior mean (red polygon). (The credible interval for P53 MUT is omitted to reduce graph clutter.) The graph show that with loss of P53 function, BRCA1 is more strongly localized to the nucleus, and fewer cells stain negatively for BRCA1.

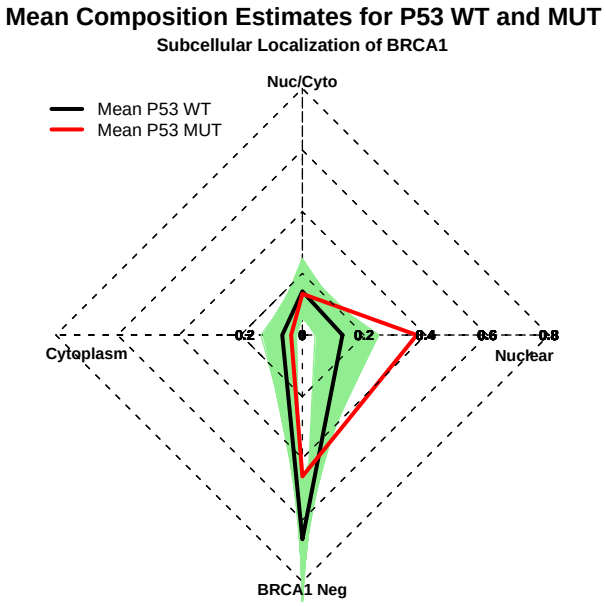

Figure 6: Mean compositions for P53 WT and P53 MUT groups (4-part BRCA1 localization). The green region indicates a 95% credible region for the P53 WT mean (the credible region for P53 MUT is omitted to reduce clutter). The mean P53 WT composition shows a large component of BRCA1 negative cells. Conversely, with the mutation of P53 we see a substantial increase of BRCA1 localization in the nucleus, and evidence of reduction in BRCA1 negative cells.

Figure 7 shows the difference composition (i.e., inverse perturbation, Billheimer, *et al. 2001* in BRCA1 localization associated with the change from P53 WT to P53 MUT. By comparing with the *identity composition* (0.25 in all four components), we observe a dramatic increase in BRCA1 nucluear localization, and a relative decrease in other components.

Although the 95% credible interval does not exclude the identity composition for the BRCA1 negative marginal composition, there is some evidence of a change in this component. Figure 8 displays the same information in an alternative format (van den Boogart, 2005). Note that when plotted in combination with "nucleus" and separately with "cytoplasm", the BRCA1 negative credible interval excludes the identity composition. This helps to clarify that BRCA1 is found in more cells with the loss of P53 WT.

We also show the posterior point estimates for each patient's BRCA1 localization mean. These values are shown in Figures 9 and 10.

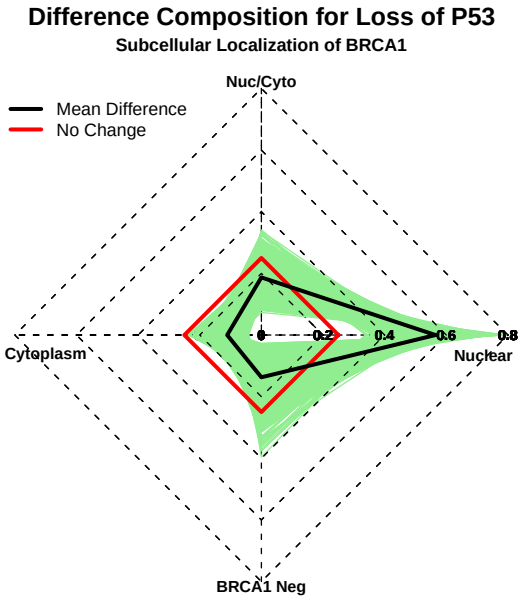

Figure 7: Mean difference composition for loss of P53 WT (4-part BRCA1 localization). The black polygon show the mean difference composition, while the green region denotes a 95% credible region for this difference. For comparison, we also show the 4-part identity composition indicating "no change" (in red). The credible region excludes the "nuclear only" and "cytoplasm only" components indicating substantial change in these components.

4.3 Association with Disease Free Survival

Finally, we wish to evaluate the association between BRCA1 subcellular localization and disease free survival (DFS). Because DFS is a patient specific outcome, we summarize each patient's BRCA1 distribution using the posterior mean point estimates above. Note that this estimate combines information from both P53 WT and P53 MUT cells, and provides a composite description of the cells within each patient. One approach to evaluating the association with DFS is to use the 4-part compositions as explanatory variables in a Cox proportional hazards regression model. This approach is easy to implement, but ignores variability in estimating patient specific compositions, and assumes them to be known. For the 31 breast cancer patients we have a median follow-up of slightly greater than 7.5 years, and observe 12 cancer recurrences.

The use of compositions as explanatory variables in linear models has been developed in the statistical field of experiments with mixtures (see for example Cornell, 2002). The standard assumption is that the expected response, $\eta(\mathbf{z})$ (with $\mathbf{z} \in \nabla^d$), depends only on the composition, $\mathbf{z}$. Different modeling choices then reduce to different forms for $\eta(\mathbf{z})$. Many arguments supporting these different choices are motivated by improving interpretation of the model parameters; the primary difficulty arising from the sum constraint of the component proportions ($\sum_i z_i = 1$).

Following Aitchison and Bacon-Shone (1984), we use the additive log-ratio transformation (alr, $\phi(\cdot)$) to transform the constrained 4-part BRCA1 compositions to unconstrained $\Re^3$. The advantage of this parameterization is that it provides direct hypothesis tests for *inactivity* of components.

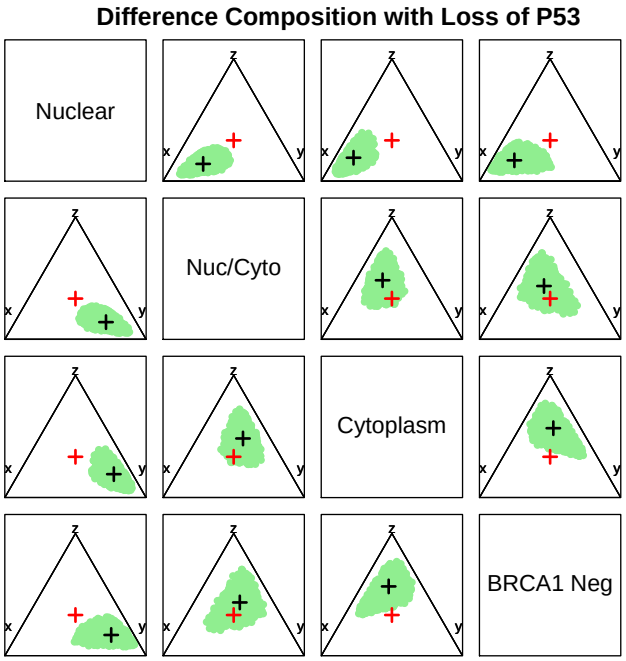

Figure 8: Mean difference composition for loss of P53 WT (4-part BRCA1 localization). The black plotting symbols show the mean difference composition, while the green region denotes a 95% credible region for this difference. For comparison, we also show the 4-part identity composition indicating "no change" (in red).

That is, tests for identifying which components have no influence on the response. Let

$$\mathbf{x} = \phi(\mathbf{z})$$

Then linear and quadratic models for the 4 component mixtures become

$$\eta_1(\mathbf{z}) = \beta_0 + \sum_{i=1}^{d} \beta_i x_i \ , \quad \eta_2(\mathbf{z}) = \eta_1(\mathbf{z}) + \sum_{i=1}^{d} \sum_{j=1}^{d} \gamma_{ij} x_i x_j$$

Becasue of the limited number of events, it is likely not appropriate to propose too rich a model structure. Indeed Harrell (2001, p.61), recommends no more than one covariate for every 10–20 events (e.g., recurrences) to avoid overfitting.

The limited biological knowledge avaiable suggests that increased nuclear BRCA1 would indicate worse prognosis. However, Cox PH regression with x_1 ($\log(z_1/z_4)$) shows no evidence of a relationship (p-value of 0.97). Similarly, regression against the linear structure ($\eta_1(\mathbf{z})$) finds little structure with any of the BRCA1 components ($p = 0.41$). Finally, (grasping at straws), we also regress DFS agains the first principal component of the patient specific compositions (Pawlowsky–Glahn and Mateu-Figueras, 2005; van den Boogaart, 2005). Again, we find no evidence of a relationship between DFS and subcellular BRCA1 localization ($p = 0.90$).

4.4 Summary and Open Questions

In a clinical data set of 31 resected breast cancer tumors we see a strong relationship between P53 status and BRCA1 subcellular localization in cancerous epithelial cells. Intact P53 is associated

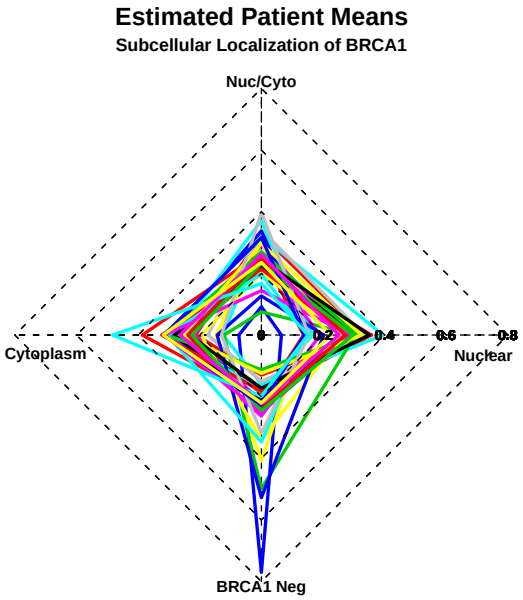

Figure 9: Posterior mean point estimates for 4-part BRCA1 localization for all 31 patients. We see that patients exhibit a range of values for each of the four subcellular components.

with a large proportion of BRCA1 negative staining cells (66%), with only 13% of cells exhibing "nuclear only" staining. Conversely, with the loss of P53 function we observe a substantial increase in nulcear only staining (36%), and a concommitant decrease in BRCA1 negative cells (46%). In this small clinical data set we can find no evidence that BRCA1 localization is associated with DFS. However, because of the small sample size, and limited number of recurrences we have little power in detecting such a relationship.

Questions

1. A parallel analysis using only the 3 part composition (nucleus, nuc/cyto, cytoplasm) similarly shows an increase in nuclear localization, and a decrease in cytosol localization of the BRCA1 protein. Similarly, we might have considered all eight P53 x BRCA1 location combinations as our categories. How should we choose the dimensionality of the categorical quantity? (i.e., Upon which sum do we condition?)

2. How should we parameterize (transform) compositional covariates? Can we gain insight from the orthonormal bases of Egozcue and others (2003)?

3. How should we use the uncertainty in patient specific posterior distributions for DFS covariates? (Obviously, this is a more general question for Bayesian inference.)

4. How should we perform model selection with compositional covariates? Is this fundamentally different than the non-constrained case?

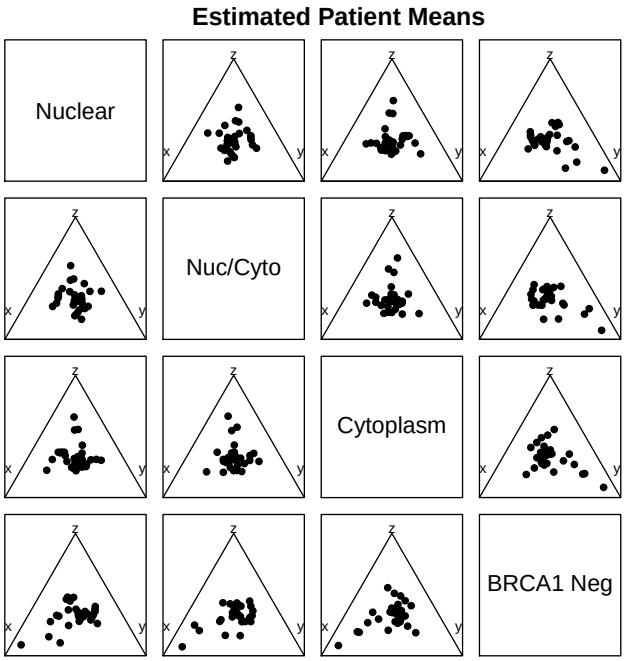

Figure 10: Alternative view of the point estimates for 4-part BRCA1 localization for all 31 patients (vand den Boogart, 2005). Patients exhibit a range of values for each of the four subcellular components.

References

Aitchison, J. (1982). The Statistical Analysis of Compositional Data (with Discussion). *J. R. Statist Soc. B 44*, pp. 139–177.

Aitchison, J., and Bacon–Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika 71*(2), pp. 323–330.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Aitchison, J. (1992). On Criteria for Measures of Compositional Difference. *Math. Geol. 24*(4), pp. 365–379.

Aitchison, J. (2003). Compositional Data Analysis: Where are we an where should we be heading? invited address *CoDaWork'03* Girona, October 2003.

Baggerly K.A., Morris J.S., Coombes K.R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics 20*(5) pp. 777–785.

Billheimer, D. (2001). Compositional Receptor Modeling. *Environmetrics 12*, pp. 451–467.

Billheimer, D., Guttorp, P. and Fagan W.F. (2001). Statistical Interpretation of Species Composition, *J. Amer. Statist. Assoc. 96*,(456) pp. 1205–14.

Cornell, J. (2002). *Experiments with Mixtures, 3rd ed.*. New York: Wiley.

Egozcue, J.J., Pawlowsky–Glahn, V., Mateu–Figueras, G., Barcelo–Vidal, C. (2003). Isometric

Logration Transformations for Compositional Data Analysis. *Math. Geol. 35*(3), pp. 279–300.

Feng, Z., Kachnic, L., zhang, J., Powell, S., Xia, F. (2004). DNA Damage Induces P53–dependent BRCA1 Nuclear Export. *J. Biol. Chem. 279*(27), pp. 28574–584.

Harrell, F. (2001). *Regression Modeling Strategies*. New York: Springer.

Karas M., Hillenkamp F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem. 60*(20), pp. 2299–2301.

Neyman, J. and Scott E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*, pp. 1–32.

Pawlowsky-Glahn, V. and J.J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *SERRA 15*(5), pp. 384-398.

Pawlowsky–Glahn, V., and Mateu–Figueras, G. (2005). The Statistical Analysis on Coordinates in Constrained Spaces, abstract in *International Statistical Institute,* 55^{th} *Session* Sydney, N.S.W.

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet 359*(9306), pp 572–577.

Ransohoff D.F. (2005). Lessons from controversy: ovarian cancer screening and serum proteomics. *J. Natl. Cancer Inst.*, *97*(4), pp. 315–319.

Rowe, D. (2003). *Multivariate Bayesian statistics: models for source separation and signal unmixing* Boca Raton, FL: Chapman and Hall/CRC.

van den Boogaart, K.G. (2005). R package: compositions, v. 0.9–10. URL *http://www.stat.boogaart.de/compositions*

Yildiz P., Shyr Y., Rahman SMJ., Wardwell NR, (14 others), Massion,PP. (2005). Approaching the diagnosis of lung cancer with serum proteomic profiling. *submitted – in revision.*

Zimmerman, L., Rahman, J., Yildiz, P., Werneke, G., Shyr, Y., Carbone, D., Liebler, D., Caprioli, R., Massion P.P. (2005) Truncation forms of serum amyloid A contribute to a serum proteomic signature of lung cancer paper presented at *American Association of Cancer Research Annual Meeting.*