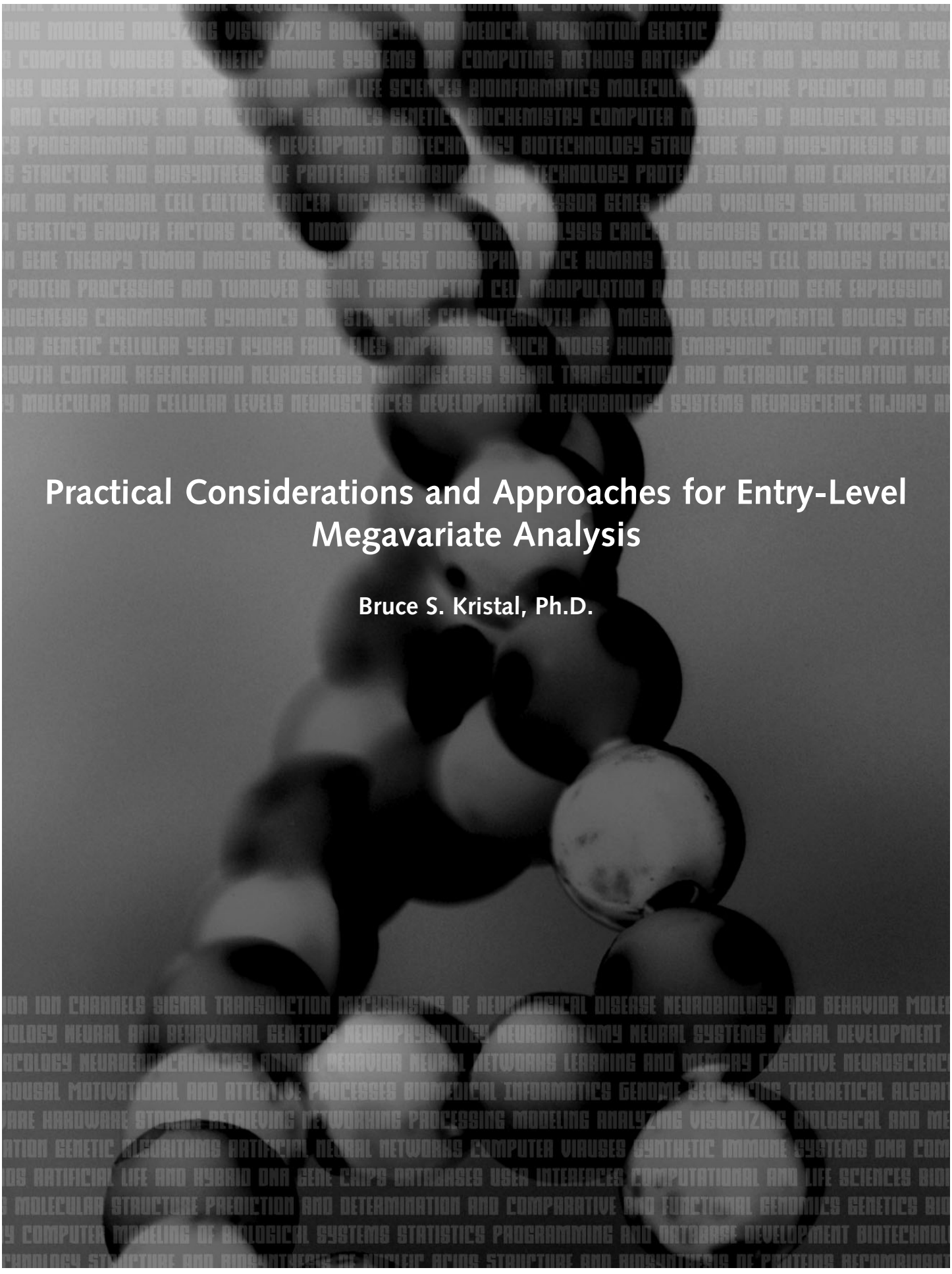


Practical Considerations and Approaches for Entry-Level Megavariate Analysis

Bruce S. Kristal, Ph.D.

OVERVIEW

High data density data acquisition techniques such as mRNA arrays, 2D-electrophoresis-based proteomics, and metabolomics studies by HPLC, NMR, or mass spectroscopy yield a wealth of data about each and every sample. Resulting datasets can readily exceed 106 datapoints, well beyond the reach of ready comprehension by most individuals. For these types of datasets, data-driven analysis can supplement – or sometimes even temporarily replace – hypothesis-driven research. An introduction to the practical aspects of three approaches to megavariate data analysis will be presented: 1) Unsupervised hierarchical cluster analysis (HCA) – grouping observations or variables without preconceived divisions. 2) Unsupervised principal components analysis (PCA) – identifying linear combinations of variables that account for the maximal portion of variation within a dataset. 3) Pattern recognition – mathematically training algorithms (often using supervised HCA or PCA) to determine the group(s) of origin of unknown samples.

The Problem: A relatively "small" dataset, 43 samples, 104 variables (4472 entries)

A SEVEN STEP APPROACH

Megavariate analysis[1] may be said to have one "goal" – to represent very large data sets in a form that humans can comprehend. In practice, this boils down to a two-part practical objective – to identify "valid" analyses (ie, to avoid making subtle mistakes with large consequences) and to identify the most powerful analysis or analyses (ie, to glean all possible useful information from a dataset). The major barrier to succeeding in these goals and objectives is that megavariate analysis is at least as much "Art" as "Science." It is, for example, not uncommon for two individuals that analyze a single data set to find some differences in the results of their analysis. In some cases, both solutions may be essentially identical (differing only in specifics), or they may be substantially different. The differences may be such that one solution is correct and the other incorrect, or the different solutions may both be correct. Likewise, even two completely valid analyses can differ substantially in the information gained. Differences can arise for many reasons. Slight differences in preprocessing or grouping methods chosen, for example, can often lead to different results. Often, seemingly small changes made far upstream in analysis can become the dominant determinants of the subsequent analysis. Not surprisingly, the more complex and/or subtle the effects being studied, the more such confounds become serious.

As such considerations hopefully make clear, we cannot hope to give a comprehensive "How-to" guide to megavariate analysis. Rather, I will attempt to give a broad map that I hope will enable one to begin to look at data using two basic exploratory approaches, HCA and PCA. I will then conclude by presenting how these exploratory approaches can be used to generate profiles capable of distinguishing groups of interest. For simplicity, we will present the initial stages of megavariate analysis in the context of eight steps:

[1]The terms "megavariate" and "multivariate" are often used interchangeably to refer to the approaches being used (eg., PCA, HCA, pattern recognition), with some individuals using one term or the other primarily to reflect the dataset size. Unfortunately, the term multivariate is also used, by others, to refer to techniques such as ANOVA and multiple regression that involve measurement of more than one variable. Therefore, for clarity, we will use the term "megavariate" throughout this text to refer to the set of techniques including HCA, PCA, and pattern recognition.

Step I: Decide Why You Are Here...

Analytical approaches designed for megavariate dataset analysis should not be considered simply as extensions of more common statistical approaches, such as t-tests, ANOVA, linear, logistic and multiple regressions, etc. While megavariate techniques can be used to extend the reach of such techniques, the primary advantages of megavariate techniques are the fundamentally different insights one can gain into data. These insights include the ability:

- To discover naturally occurring (often unexpected) groupings in your data.
- To discern discontinuous relationships that may occur in your data.
- To discern relationships between variables.
- To work with datasets having substantial levels of missing data.
- To identify variables that distinguish two or more classes.
- To classify unknown observations into known classes.
- To build mathematical models of large datasets.
- To compress large datasets into smaller, more informative datasets
- Etc....

These advantages do, however, come with a series of disadvantages. These include:

- The relatively steep learning curve for some aspects of this work.
- The potential to make tremendous mistakes with little notice.
- Financial costs associated with software acquisition.

These disadvantages can often be readily circumvented, however, bringing the power of megavariate analysis readily within the realm of feasibility. Aiding this effort is the advent, especially within the last 5-10 years, of a broad series of software packages that enable the practical aspects of analysis to be conducted readily. Indeed, probably anyone with a solid high school training in math and science could master the programs themselves. The subtleties of analysis are, however, clearly present a different problem.

The first question, therefore, is "Why are you here?". A researcher considering these (megavariate) approaches has to determine the EXACT question of interest. At first glance, this is trivial, but even a little experience with multivariate analysis readily drives home the message of how critical subtle differences in the question fundamentally alter both approaches and results. The question of taking a series of samples from, for example, Alzheimer patient and control samples can be very informative, but it is necessary to consider the question of interest. Do we care if we can distinguish classes? Do we want to know if there are differences? If they are significant (very different questions!)? Which variables change? Which variables change significantly? Whether there are overall shifts in metabolites/variables? Which variables interact? Which variable are related? Are their potential outliers (theoretically, we'll deal with "practically" later).

As one example, clustering approaches (e.g., HCA) tell you whether there are naturally occurring groups in the data and, if so, what the program feels the dominant grouping is. It therefore seems reasonable to take, for example, a mixture of Alzheimer's and control patients, and ask whether information on some series of variables is sufficient to classify them. This may work, but it is a different question than asking whether a common set of variables distinguishes female AD patients from female controls and male patients from male controls. It is likewise a different question than asking about the two sexes independently. It is likewise a different question than asking whether biochemical differences in AD patients are drug, diet or environmentally neutral. In univariate analysis, or some types of multivariate analysis, this question might be addressed by progressively paring the dataset, or perhaps through multiple or logistic regression. While these approaches might work, megavariate analysis offers complementary

approaches, such as examining localized clusters in principal component space, examining loadings on principal components, or examining subgroups within cluster analysis.

A second issue concerns variable studies. Clustering shows relationships. Univariate analysis offers little other than regression coefficients to show when two or more variables are related. While this is both clear and useful for a few variables, it is unwieldy for datasets with, perhaps, 10 variables (90 correlations), and is impractical for very large matrices (correlations across a microarray experiment could readily exceed 10,000 x 10,000 or 100,000,000 entries.

Other questions:
- Do you need statistical significance (unlikely with clustering techniques)
- Do you need to know why groups exist?
- Do you need to know why groups don't exist?
- Do you need to know whether a given variable/set of variables correlate with an outcome?
- Do you need to distinguish current groups?
- Do you need to distinguish future groups?
- Can your assays be maintained over time?
- Etc…

As a general, first approximation, single direct questions point to single megavariate approaches (HCA, PCA, PLS [partial least squares], PLS-DA [partial least squares projection to latent structures-discriminant analysis], Machine learning, etc). Combinations of such questions and questions that can not be precisely limited often require the investigator into either the use of multiple data analysis approaches/packages and/or will considerably limit availability of packages in a non-sequential and non-linear way. Why and why not questions often require sequential nested models within one or more analyses.

Step II: Find Help!!!

Once you have determined the exact question of interest, the next step is to FIND HELP. At least periodic contact and discussions with people whose specialty is in the analysis of large datasets is nearly essential to reaching optimal solutions to problems in this area without prohibitive investments in time, money, and aggravation. These people might be collaborators, department members, consultants, or even the software developers (definitely consider your source!). The suggestion to look for help comes for several reasons. Some have to do with the theoretical complexity of the field, some with the subtleties involved in addressing some questions, some with the nomenclature, some with the practical choice of software, and some with the practical complexity of the software analysis (the analysis itself is trivial, the options can be daunting – and critical).

Megavariate analysis, at least as it applies to bioinformatics, computational biology, and mathematical modeling of biological processes, is an evolving field. With the exception of the most basic terms, the terminology and notation is almost always specific to the software package used. Many available packages were not designed with biologists in mind, and most biologists do not have extensive training in this area. Available statistical packages available often have very discrete and non-overlapping strengths and weaknesses. Many have no over-fitting controls. As multivariate and pattern recognition algorithms are often designed, essentially, to always find an answer (even if one doesn't exist), data can readily "walk you off a cliff" (eg., over-fitting models) with no warning. Note that there are times that this may not be a problem. Understanding when it is goes back to point #1 – What is your EXACT question. For some questions, walking off a cliff is, indeed, not a problem (e.g., if all you care about is describing your data, over-fitting may not matter)… (but be careful when you decide this)

Two final points:

- Most mistakes that can be made will be readily avoided by minimal contact with an expert
- Under some cases, having the expert do your analysis, if feasible, may be your best solution

Step III: Examine Available Resources

Asking around can often save substantial amounts of time and money. Commercial software packages capable of megavariate analysis range from a few hundred dollars to over one hundred thousand dollars. In addition, some informatics problems require custom programming. Some problems require access to specialized databases that are themselves costly (e.g., genome databases). Secondly, while getting started with most programs is very quick, utilizing the programs potential often requires both a relatively steep learning curve and substantial practice and experience to develop a feel for the program and its strengths and weaknesses. Again, these problems can often be readily avoided.

Many universities today have licenses (either complete site licenses or at least discounted subscriber licenses that can reduce your costs substantially). Even if your university does not offer software resources, two other sources should not be overlooked. First, many of the companies that sell commercial packages have available, fully or largely functional demos. This is often enough to determine if a given package will work for your needs. As we will discuss in more detail below, packages tend to have very specific advantages and disadvantages, and they should be carefully evaluated prior to purchase. **DON'T ASSUME MEGAVARIATE ANALYSIS SOFTWARE PACKAGES ARE INTERCHANGEABLE!!** Unlike, for example, word processing programs or basic statistics programs, megavariate data analysis programs often offer completely different technologies. The other consideration is to discuss your needs with the people who you have gotten to help you in Step 1: **DO THIS FIRST. DON'T WAIT!**

Perhaps more importantly, initial looks at even extremely complex datasets by some approaches can often be carried out in less than an hour, often within 5 minutes. Such a feasibility study can save you both financially and, perhaps more importantly, in time required. It is often possible, with even such a rapid look, to determine if the subsequent investment in time and resources needed to fully enter this area are worthwhile.

Step IV: I'm Committed, Now What: Five Basic Questions

While many different factors will eventually weigh on the final choice of megavariate analysis approach taken, at least five issues must be considered up front: The EXACT question of interest, the nature of the data, your willingness to conduct the analysis objectively or subjectively, the structure of the dataset, and the ability to repeat experiments/are your datasets independent.

The first of these, the EXACT question of interest, was also dealt with above. Now, however, is an important time to reconsider your answer to this question. When we first examined this question, it was essentially as a neophyte. By the time you have "arrived" at this step, you will have talked with people experienced with multivariate analysis, and maybe looked a little at what these programs can do with your data. Thus, before really beginning a detailed study, it is worth re-evaluating this question to see whether any or all aspects of what you previously thought was important has changed.

The second question deals with the nature of the data and data handling. In a simple univariate experiment, one might have, for example, a series of enzyme measurements from a study of two groups. Such a dataset appears logically set up for a t-test. Likewise, the investigator might move to paired t-tests if the experimental design was appropriate, or to ANOVA, for multiple comparisons. If the investigator is careful, or statistically astute, they might precede this test with tests for normality and homoscedasticity . Likewise, investigators might choose to remove statistical outliers (hopefully following some consistent protocol!), or manipulate the data for subsequent analysis (e.g.,, ratio, log transform, normalizations, etc). Such issues are also appropriate for multivariate analysis. But multivariate analysis also raises other issues, such as the handling of missing data (example, some programs can conduct studies where 99% of the data cells are vacant), the handling of both sample and variable outliers (replace with blank, replace with a set value, etc…), and the handling of non-numeric data. **Note that the requirement for progressively more powerful and/or complex data manipulation will considerably limit availability of packages in a non-sequential and non-linear way. Alternatively, one may need to use multiple programs.**

Another aspect of the second question is: "Are there Y variables, outcomes or classes? Age at death, disease vs. control time to loss of function, etc. The presence of a Y variable enables one to consider a set of approaches based on mapping the X-block (variables) onto the Y block (outcomes), such as PLS. Alternatively, the "outcome" may be class membership (e.g., disease A vs. B). This opens one to pattern recognition techniques such as KNN, SIMCA, and PLS-DA (discriminant analysis) [see below]. Alternatively, in the absence of such an outcome or Y variable, one chooses instead to focus on exploratory and analysis and learning more about your data by HCA and PCA. The answer to this question is neither good nor bad, again it simply helps point you in a direction.

The third question deals with the structure of the dataset. Datasets can be sample poor and variable poor (SPVP), sample rich and variable poor (SRVP), sample poor and variable rich (SPVR), or rich in both samples and variables (SRVR). Overall, megavariate analysis is not particularly useful on SPVP datasets. This is in part because of over-fitting and in part because little is often lost by using more standard approaches. SRVP datasets can generally be analyzed by most megavariate algorithms. In particular, clustering algorithms work well by breaking the samples up into classes that may be best analyzed separately. In such environments, however, it is important to make sure that the variables are relatively independent, otherwise proportional over-representation in the dataset can distort the data. Note that many aspects of the analysis of SRVP datasets may be amenable to univariate analyses. SPVR datasets, such as those in most microarray experiments, offer unique limitations. The most important of these limitations is that the megavariate analysis programs can readily over-fit such models

(imagine the trivial case of 2 cancer samples, 2 control samples, and 15,000 variables). It should be readily apparent that a large number of variables could be identified that would allow the groups to be distinguished (indeed, one could find variables that would distinguish the two sets of one patient/one control from each other). Thus, one often has to consider ways of addressing such problems (hierarchical modeling on combined variables being an example). Finally, SRVR datasets allow nearly all approaches.

A further issue related to dataset structure is that certain types of algorithms are more prone to certain problems. As one example, clustering can be done using agglomerative techniques (in which groups are created from single samples) or divisive clustering, in which the single group containing all samples is progressively cleaved. Agglomerative clustering is much less sensitive to over-fitting data in SPVR datasets. Thus, the nature of your dataset may limit your options.

The fourth question deals with whether you prefer to analyze your data objectively, subjectively, or in some combination. Simply dropping your data into a cluster analysis program and publishing the output is nice, simple, and is perhaps the purest form of analysis, but it is rarely optimal. Indeed, although megavariate data analysis can be conducted in a largely objective way, it is often helped by making critical subjective decisions. For example, the decision to remove outliers is, at some level, a subjective rather than objective decision. Preprocessing (scaling, transforming) data is often essential, but the approaches chosen can be subjective and often have substantial effects on downstream analysis. Data can be examined with each variable given equal weight, or variables can be combined into hierarchical variables. As an example of the latter, a study of gene expression within different regions of the brain can be grouped by brain vs. liver, cerebrum vs. cerebellum, or by, for example, specific layers of the cortex or parts of other brain sub-regions. In each case, the decisions made impact the final outcome. A study examining actions of the fore vs. hindbrain might be well served by a "two-compartment" model of the brain, whereas a search to explore the overall workings might well require a model with, for example, the brain divided into 20 or more regions. While these issues are, conceptually, no different than equivalent aspects of univariate analyses, the impact of such changes may be less transparent to most users.

At another level, subjectivity can be used to conduct a primary optimization. For examples, in our studies on dietary influence on the metabolome, we first examined the data objectively in a subjective way, that is, we used our knowledge of group identity to conduct an objective study. This primary study enabled us to demonstrate principle, that diet was reflected in sera metabolites. We then needed secondary studies, conducted strictly objectively to show that the markers determined were objectively valid.

As yet another example, multiple algorithms with slight differences can be used to conduct exploratory analysis (more below), choosing between these can be objective, eg., each of six methods classifies these x samples the same, so we will accept that and not further consider those whose classification is algorithm dependent. Alternatively, one may look subjectively first, determine the basic approaches that appear most promising, then again come back to the main problem. Finally, it is worth noting that initially objective studies often become subjective when one looks at the initial data and says, "No…" For example, nearest neighbor grouping algorithms fail to distinguish groups in our metabolome data even when other algorithms have >90% accuracy. This was eventually shown to result from the distribution of metabolites within the principal component space, but was originally determined more by the failure of this mathematical approach to look like any other. **Overall, objectivity adds "believability" or "credibility," whereas subjectivity adds power. These best of all worlds is when both can be used.**

The fifth question, arguably the simplest, is whether or not you can repeat an experiment. Relatively cheap, simple experiments on bacterial cultures is an obvious example of where one can readily repeat a series as often as wanted. Microarray experiments on brain biopsies of patients with rare conditions (if such studies exist) is an obvious example where repeats are impractical or impossible. Most experiments fall in the middle of the continuum implied by this question. The ability to repeat an experiment has three implications: 1) One can examine biological variability, 2) One can examine analytical variability, 3) Different techniques become available. When one conducts an experiment on a single variable in two groups, the odds that that variable will appear statistically significant at $p<0.05$ is, by definition, 5%. When one runs a microarray experiment with 15,000 variables, one needs to expect that 750 genes will appeared altered purely by statistical fluctuation, again, by definition (Note: in practice, one may reduce the effective number by cutting the dataset based on minimal expression or other approaches). If one repeats the experiment again, on only these 750 genes, one must still expect ~38 genes to appear statistically different. This both highlights the extreme difficulties of univariate analysis on large datasets as well as the advantages gained by megavariate approaches. If one now moves to analytical variability, the same arguments hold, that 5% of genes will be altered in a statistically different way than the others. While the exact statistical concerns are determined by the assays coefficient of variation (CV), one must recognize this problem. For example, if one conducts a series of enzyme assays with CVs of <2% in a study whose populations differ by 50%, then analytical variability is irrelevant. On the other hand, if one tries to conduct a study of populations that differ by 50% using a technique with a CV of 25%, the extreme analytical outliers will result in a large number of false positives, as well as false negatives. The false negative issue also confounds the replicate strategy (consider an experiment with a power of 0.8 and 100 true positives, 80 survive the first round [vs. 750 false positives] and 64 the second round [vs. 38 false positives]. Thus, even after two rounds of microarray experiments, ~33% of the data is still wrong, without considering analytical issues]). As is probably apparent, the aforementioned concerns also apply to univariate analysis, they are just exacerbated as the number of variables increases. Note, however, that megavariate analysis also adds another level of complexity. Specifically, many techniques and approaches can only be used if you have multiple data sets. One example is the subjective/objective discrimination described above. A second is the valid use of any trained or supervised algorithms. Thus, the ability to repeat an experiment puts substantial constraints on what can be done about false positives and negatives (and hence the conclusions that can be drawn) and upon the algorithms that can be used.

Thus at this point, the researcher returns full circle, to – "What is the exact question." Passage through the six questions above will often lead the researcher to a relatively narrow set of options. Thus, "What is the exact question?" reappears, as the initial questions may have been rendered impossible (eg, by inability to repeat, or by the structure of the dataset). Alternatively, understanding of the above questions may have left the researcher with specific questions not previously considered. For example, the choice of grouping algorithm in cluster analysis can be chosen to define groups based on large or small distinctions in the variable sets used. Preprocessing techniques can be used to further refine specifics (do we want to examine separations under conditions scaled to unit variance, or do you want to over [or under] emphasize variables with greater variance). These more subtle questions will be addressed below.

Step V: Now That I'm Ready, Where Do I Start: Data Preparation

Data preparation involves four stages: Input, scaling, transformations, and outlier diagnostics.
At least in the programs with which I am familiar, data are input directly from other spreadsheet type programs, such as Excel. Because most megavariate analysis programs have comparatively limited data manipulation tools, it is generally best to optimize your dataset in the other program. The final format of your data should be in will be dictated by the program you are using. When possible, it is often very helpful to put in a series of additional columns/rows that serve to describe your data. For example, class membership (where known), date of analysis, M/F, diagnosis, etc. In programs that allow this data to be present without interfering, these aid in coding samples and following analysis. Names should be chosen to be as short yet informative as possible.

Data are scaled so that each variable has an approximately equal influence (mathematically) on the subsequent analysis. As a counter example, if one groups based on height and weight without scaling, then the choice of units (eg, millimeters vs. centimeters vs. meters vs. kilometers and milligrams vs. grams vs. kilograms) becomes dominant to the actual importance of height vs. weight in establishing groups. The most common scaling approach, mean centering followed by variance scaling, is called by many names (eg, Autoscale, unit-variance scaling). Other approaches include mean-centering alone, variance scaling alone, range scaling, etc. Note that the choice of scaling can dominate all subsequent analysis decisions. Make sure you understand what the scaling is doing (eg, for a microarray experiment, do you want to over-weight or underweight results from genes that differ substantially within the dataset?)

Data are transformed to help with at least two problems. First, studies on non-transformed data can often be disproportionately skewed by data on one end of the distribution. This is often a major issue when the min-max ratio in the dataset is <0.1 - 0.01. Log transforms are commonly used to solve these and other distribution problems. Transformations can also be used to account for changes in sensitivity across experiments. One can normalize, for example, to the total signal in a study.

In univariate analysis, eg t-tests, one outlier can often hide statistical significance. Outliers are also fairly straight-forward, and are usually simply removed. In megavariate analysis, outlier analysis involves several distinct issues and often multiple steps. In the initial phase, one can examine each variable's data for outliers using univariate statistical approaches. Outliers may be removed, and blanks inserted, or the data may be winsorized -- blocks filled with a specifically defined value (eg., an upper bound, a lower bound, or a neutral value). Once one begins analysis, samples may be observed to be outliers with respect to the X-block model, the Y-block model, or with respect to overall sample residuals. Treatment of these outliers is case-specific, and often requires subjective decisions.

Step 6: Primary Analysis

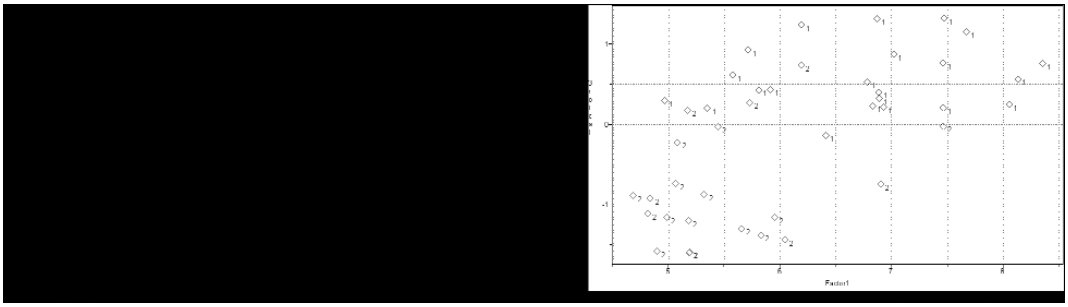
At this point, your data is ready for its first pass through the megavariate software of your choice. Overall, this stage of basic exploratory analysis is essentially trivial. Input the data, input parameters chosen above, and run the analysis. For at least the exploratory analysis programs that I have any familiarity with, this stage is very easy and fast, assuming you know the program. Even if you are using a program for the first time, it will probably take you less than a day to understand the logistics of entering data, starting analysis, and seeing results. Furthermore, once you have worked through, and begun to understand the above issues (considerably more time consuming!), those initial analyses are likely to be very informative. At this point, I would like to briefly introduce the three types of analysis. For HCA and PCA, analysis follows a common path. Data must be put into a specific format (often in Excel). Data are then input using cut and paste or program specific input protocols (usually very simple). In some programs, data is then preprocessed and transformed. In others, this manipulation occurs simultaneously with analysis. Data analysis methods and parameters are chosen, and the analysis is run. Even relatively large datasets can usually be analyzed in seconds with these approaches. Now, we will present an introduction to the specifics of PCA and HCA.

Hierarchical Cluster Analysis (HCA) is a method of data analysis that emphasizes the natural groupings of the dataset. In contrast to analytical methods that emphasize distinguishing differences between two groups, HCA uses algorithms that reduce complex data sets to establish these groups without preconceived divisions. Some programs (eg, Genelinker Gold) support simultaneous clustering of samples and variables. Clustering programs may be hierarchical, ie, beginning with a single group and showing all subsequent branches, but other clustering approaches also exist, such as those that focus on building/identifying lots of small clusters in the data, without examining their place in a larger context. Clustering approaches tend to be rigorously objective after the start of analysis, which can be either an advantage or a disadvantage. Note also that clustering can be very sensitive to aspects of your data that are not of interest (eg., in comparing AD and control differences, differences between males and females, or the clinic that processed the samples, may be dominant to disease markers)

Clustering approaches clearly have utility in some data sets (consider evolutionary analysis as one example). We have been successful in clustering one dataset based on the relationships between genotype and longevity. In contrast, our experience in studying metabolomics has shown that clustering techniques, such as HCA, are less successful in these studies, an observation that appears to result from asymmetric distributions of the samples of interest. This may be because clustering is largely a function of the overall distance between two or more variables/samples, without necessarily capturing "direction." In contrast, we have had success with component-based (also called projection-based) analyses, such as principal components analysis.
HCA (Left) and PCA (Right) simplify the data shown above. The numbers 1 and 2 refer to groups.

Principal Component Analysis (PCA), also called Eigenvector analysis, is used to determine linear combinations of original metabolites that account for maximal variation. Thus, PCA can be used to reduce the dimensionality of the data by using only some of the Eigenvectors. Lower number principal components possess greater ability to explain variation in the dataset, i.e. the ability of principal component 1 to explain variation is greater than that of principal component 2, etc. For our purposes (ie, classification), the subset of Eigenvectors chosen can then be evaluated in terms of their ability to distinguish members of the different groups. More important, in the context of our long range goals, PCA can be used to determine which of the multiple compounds that may differ between humans with different conditions are the most useful for classification and drug development purposes.

HCA (Left) and PCA (Right simplify te data shown above. The numbers 1 and 2 refer to groups.



The great strength of PCA is that it gives a rapid, often very powerful view of your data, simplified and graphed in two or three dimensions. Initial looks at PCA plots reveal groups in the data. For example, color coding samples by sex or by group of interest (available in most programs) gives a rapid insight into whether such issues affect grouping within your set. Coding by analytical number or by order of data collection often reveals analytical drift (a major issue with complex analytical techniques). When samples of interest that "should" be grouped are, in fact, distinct, PCA gives warning to a problem in understanding or previously hidden complexity within the data. Further looks at how each variable contributes to the separation in each dimension (ie, on each component) gives insights into the relative importance of each variable.

Pattern recognition-based approaches take these (and other) exploratory analyses a step further. Pattern recognition has three essential stages. In stage 1, a set of samples are used to "train" a program/algorithm to recognize members of specific classes (e.g., AD patients/controls) using a defined set of variables and criteria (e.g., 15 specific variables, all mean-centered, variance scaled, log transformed, and winsorized at 3 SD). The specific criteria used are dependent on the pattern recognition algorithm, of which there are probably 100's to 1000's. In stage 2, the trained algorithms are presented a set of unknowns, and the algorithms "attempt" to classify them. In stage 3, failures (and sometimes successes) are assessed (sometimes manually, sometimes by automated techniques, such as neural net-based algorithms), and the algorithms "improved." The process can continue iteratively, often 1000's of times for some machine based learning protocols. Note that over-fitting issues are critical, and it is important to retain datasets that are not used for training and model improvement for final testing. Overall, pattern recognition algorithms can improve the quality of the model (and test it) by training on subsets of the data (bootstrapping), can determine those components of the model (i.e., specific sera constituents) that are most important in determining the models' character, and can be used even when some data are missing or highly unusual in a specific sample (e.g., a constituent is not detected).

The critical issue/advantage of pattern recognition is its ability to provide class prediction, and associated with this, its ability to identify variables important for class prediction. This stands in direct comparison with HCA and PCA, both of which give insight into your data, but which make no specific claims or predictions. The two classification algorithms I'll discuss briefly are K-Nearest Neighbor (KNN) and Soft Independent Modeling of

Class Analogy (SIMCA). Of the programs with which I am familiar, both are available in the data analysis package Pirouette; SIMCA is also available in the SIMCA-P9 package (see below). KNN is based on the same mathematical theory as HCA (KNN at k =1 is HCA) and constructs models where the k nearest neighbors "vote" for class membership. KNN is a distance based metric with advantages in tight, uniformly-distributed datasets. SIMCA is based on the same mathematical theory as PCA (ie, SIMCA is component based). While PCA calculates principal components on a whole data set, SIMCA provides principal component models for each class within the training set. SIMCA's great strength is enabling one to distinguish complex groups, and in gaining better insight into the structure of a class. Note that KNN will always assign a class to an unknown, whereas SIMCA may assign an unknown to one class, multiple classes, or no known class. Both SIMCA and KNN are based on the assumption that the closer samples lie in measurement space, the more likely they belong to the same category. Alternative distribution models (separability, probability) are also available, but both tend to over-fit sample-poor datasets.

Step VII: At First Look My Data Are: A Potentially Iinteresting; B) Completely Uninteresting; C) Completely Indecipherable; D) Awe-Ispiring — Now What?

As noted above, even initial exploratory looks at data are often enough to give fairly strong insight into the future of this line of analysis.

If your data are POTENTIALLY INTERESTING, the first stage is to identify what aspects of the data make it less interesting, and see why they are present. For example, if you are trying to group observations into classes, and you have 95% accuracy, look at the outliers. Are they general outliers or is one or more variables particularly abnormal for a given class? Are the samples outliers biologically or analytically (going back to your lab notebooks and analytical notebooks can often explain outliers)? If these don't help, revisit alternative transforms, scales, grouping algorithms, etc. Note that, depending on your conclusions and your system, you need to decide carefully how this "retrofitting" meets scientific standards and how to report what you have done. For example, it is probably unnecessary to report that your data enabled you to correct a technician's typing error, but it is also unethical to find the one analysis that worked and report it as if it was the only one tried.

The second stage of analysis when dealing with "POTENTIALLY INTERESTING" data is to identify the aspects of the data that make it more interesting. For example, seeing (by elimination, if necessary), which variables are useful as class predictors. Sometimes this can be great, yielding information of variables not previously known to be important (although be careful of overfitting), and sometimes this can be deflating (eg, discovering that males and females have biochemical differences). Once additional work has clarified these findings, it is useful to look for the simplest way of presenting and clarifying your work for public consumption, which retaining the information needed to reconstruct analysis, if necessary.

COMPLETELY UNINTERESTING (ie, "negative" data) and COMPLETELY UNINTERPRETABLE (ie confusing) results are, of course, fundamentally different, but they can be addressed through essentially the same process. 1) check all data entry and analysis steps, make sure you are actually running the analysis that you think you are, on the data set that you think you are working on. Transformation and scaling issues, in particular, may be problems. 2) Retry the analysis with different parameters, sometimes one approach will fail, whereas several related approaches may succeed. 3) Retry with a different basic analysis, for example move from clusters to principal components. 4) Look at variable inter-relationships, maybe you have multiple variables present that overlap (high correlation coefficients), and this is skewing your dataset. 5) Rethink the initial question. Maybe one needs to reconsider having, for example, both males and females in the same dataset. 6) Reconsider experimental design, for example, do you have too many variables and too few samples? 7) Consider alternative hypotheses. 8) If after all this, the dataset still looks negative, consider what a negative result would mean, and perhaps use this to guide future lines of investigation. Megavariate data sets have been, for example, very effective in showing us flaws in our reasoning on mitochondrial involvement in AD.

AWE-INSPRING results are actually fairly common in megavariate analysis. In part, this is because these techniques do indeed provide tremendous insight into complex issues in biochemistry; unfortunately, it is also true, at least in part, because these algorithms can overfit data and can find and exploit even subtle experimental flaws. Thus, awe-inspiring results should be checked in at least four ways: 1) Make sure the answer is robust (if it is that awe-inspiring, the result should be robust across, for example, different grouping algorithms). 2) Make sure the data was entered equivalently and correctly 3) Check experimental design (eg., one problem we've seen is in a data set where the controls and experimental samples were differentially present from different cohorts, and the processing, supposedly identical across the groups, was sufficient to alter analysis. 4) Check for over-fit, either with internal diagnostics or with cross-validation approaches, such as randomly leaving out at least two observations at a time.

Step VIII: Where Do I Go Next?

The steps above should provide a rough road map through the earliest stages of entry into megavariate analysis. At this point, however, the roads diverge as specialized interests relating to specific problems, specific datasets and models become the predominant issues. These issues are beyond the scope of this presentation.

SOME PROGRAMS OF INTEREST

Cautionary Note: My megavariate work focuses nearly exclusively on classification analysis of megavariate metabolomics data related to nutritional epidemiology studies. These studies include both rats and humans. The nature of these studies and my analytical requirements has limited the programs I have looked at and worked with. In addition, I was led to several programs by collaborators, advisors, and some presentations I have seen at meetings, rather than by an extensive survey of the field. For these reasons, the two primary megavariate analysis programs that we use (Pirouette and SIMCA-P9) are likely quite different that those more commonly used by most biologists. Indeed, if your focus is microarrays I would suggest that you take a serious look at programs designed for that purpose as well. Nonetheless, my thoughts and comments…

Basic Programs for data handling and statistical analysis

EXCEL, ACCESS: Used for data management, basic data manipulation
STATVIEW, NCSS, PASS: Primary univariate data analysis, normality testing, Power analysis

Megavariate Analysis programs

PIROUETTE (INFOMETRIX): Pirouette is a multivariate analysis program containing both clustering (eg HCA, KNN) and component (PCA, SIMCA) algorithms. We use this program for all clustering analysis, and some introductory component modeling, especially outlier diagnostics. The program can also examine variable modeling and fit to models. Very fast and easy to use, very good clustering algorithms. Excellent manual. Good ability to go back and forth between projection and clustering methods. Good outlier diagnostics within PCA. Weaknesses of this program include relatively limited capacity to handle missing data, lack of some advanced analyses, and lack of over-fit diagnostics. My choice for your first program.

SIMCA-P9 (UMETRICS): SIMCA-P9 is a workhouse projection-method based program, containing a constellation of features required for some projects. These include outlier and overfit diagnostics, PLS-DA (discriminant analysis) and SIMCA analyses, and ability to examine fit to model in both X- and Y- dimensions. The ability of specific variables to influence projections can also be determined. SIMCA-P9 is robust even in the presence of missing data (eg, analytes not measured due to analytical difficulties or interferences). We have used this program to show our ability to distinguish dietary groups (with >95% accuracy) and to determine those metabolite peaks that might define the distinctions between these classes of samples. Weaknesses of this program include its complexity, lack of clustering, and limited ability to save work in progress. IF YOU ARE GOING TO USE THIS PROGRAM, READ THE MANUALS AND RELATED TEXTS, AND READ THEM VERY CAREFULLY. This program appears intended for people who need powerful analysis and are serious about learning to do them correctly. My choice if you have to do complex projection analysis. (Note: SIMCA-P10 should be available by the time of Neuroscience).

GENELINKER GOLD (MOLECULAR MINING): The only one of the three designed for biologists. Has both principal component and clustering capacity. Also, unlike the others, has self-organizing maps, 2D clustering, and false color maps (ie, as array data is often presented). Relatively user friendly and powerful. I am a relatively new user and can't really comment further at the time of writing this manual. Feel free to talk with me at the meeting.

SOME BOOKS/MANUALS OF INTEREST:

Multivariate Statistical Methods; A Primer: Bryan Manley (Good, basic introductory book), ISBN 0-412-60300-4

Pirouette Manual (available online in the Pirouette demo, Good introductory text)

Solving Data Mining Problems through Pattern Recognition, Ruby Kennedy, Yuchun Lee, Benjamin van Roy, Christopher Reed, Richard Lippman (Some emphasis on a program called PRW-Pro); ISBN 0-13-095083-1

Multi- and Megavariate Data Analysis: Principles and Applications, L. Eriksson, E. Johansson, N. Kettaneh-Wold, ISBN 91-973730-1-X (Substantial emphasis on SIMCA-P