

## **Megavariate Statistics meets High Data-density Analytical Methods: The Future of Medical Diagnostics?**

David J. GRAINGER

*Department of Medicine, Box 157, Addenbrooke's Hospital, Cambridge, CB2 2QQ, U.K.*

Recent advances in high data-density analytical techniques offer unrivalled promise for improved medical diagnostics in the coming decade. Genomics, proteomics and metabonomics (as well as a whole slew of less well known “omics” technologies) provide a detailed descriptor of each individual. Relating the large quantity of data on many different individuals to their current (and possibly even future) phenotype is a task not well suited to classical multivariate statistics. The datasets generated by “omics” techniques very often violate the requirements for multiple regression. However, another statistical approach exists, which is already well established in areas such as medicinal chemistry and process control, but which is new to medical diagnostics, that can overcome these problems. This approach, called megavariate analysis (MVA), has the potential to revolutionise medical diagnostics in a broad range of diseases. It opens up the possibility of expert systems that can diagnose the presence of many different diseases simultaneously, and even make exacting predictions about the future diseases an individual is likely to suffer from.

KEY WORDS: *PCA, PLS-DA, genomics, proteomics, metabonomics, immunomics*

The classical approach to medical diagnosis involves collecting data about the individual under study in a stepwise fashion in order to eliminate possible causes for the exhibited symptoms one at a time, until the actual cause has been identified. Such differential diagnosis relies on relatively few measured variables (which might be observations, such as the presence of a rash, physiological tests, such as blood pressure or temperature, or sophisticated biochemical measures ranging from blood glucose to cardiac enzymes). Although sometimes time consuming or expensive, this approach works well for most diseases, although there remains a number of prevalent diseases where a firm diagnosis cannot be made until after death (of which the best known is Alzheimer's Disease).

Data generated in this way can also be used to assess future risk of disease. The risk factors may be the same parameters that are used to diagnose the presence of the disease, or they may be additional measurements, which have been identified from cross-sectional epidemiological studies. Either way, the resulting dataset can be analysed by classical multiple regression statistics to generate a model of future disease risk. The PROCAM dataset from Münster, Germany, is an excellent example of this approach applied to heart disease.

The dataset generated from such experimental approaches usually meets the requirements for multiple regression statistics. Firstly, the data table is usually “long and lean” – that is, we usually have data on more individuals than there are variables that we have measured on each individual. The data matrix is described as having full rank, which is a prerequisite for multiple regression (the situation is analogous to simultaneous equations – you need  $n-1$  equations to solve a problem with  $n$  variables – with less equations you can only provide an incomplete solution). Secondly, and equally importantly, the variables are usually unrelated. The source of the majority of the variation blood glucose levels, for example, is different from the source of most of the variation blood pressure, for example. Although two classical parameters may correlate to some degree, they usually do so only weakly. Again, this is consistent with the requirement of multiple regression that the variables be independent. Finally, because only a few variables have been measured the dataset is usually complete, or almost so - multiple regression is intolerant of missing data.

---

*Abbreviations used in this paper:* MVA, megavariate analysis; PCA, principle component analysis; PLS, projection to latent structures by partial least squares; PLS-DA, PLS discriminant analysis; QSAR, quantitative structure activity relationship

However, the downside of this approach is that the small numbers of variables measured on each individual do not usually describe sufficient variation to allow classification of individuals on a case-by-case basis. Essentially, they contain too little information to allow a firm diagnosis or prognosis to be made for an individual. They can provide risk stratification models, where certain individuals can be assigned a risk 10-fold or more higher than the population average, but there is usually insufficient information inherent within the dataset to allow more precise individual-by-individual diagnosis.

*“omics” technologies generate a more powerful individual descriptor*

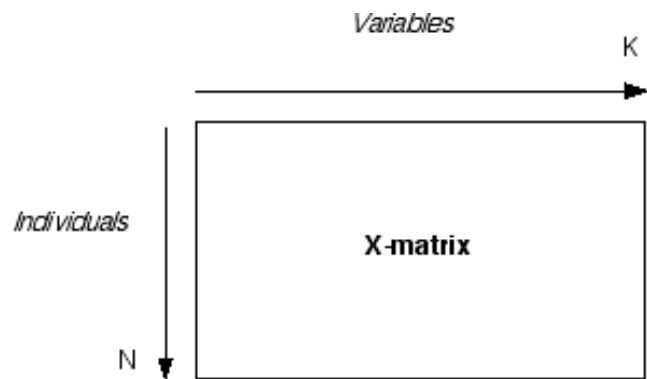
With the advent of analytical approaches capable of making hundreds or even thousands of measurements on a biological specimen simultaneously, it has now become possible to generate a considerably more detailed descriptor of the status of an organism. Such profiles contain sufficient information for everyone to have a unique descriptor – rather like a fingerprint. Such “omics” descriptors, therefore, might in principle allow individual diagnosis and prognosis to be made with a medically useful degree of accuracy.

Genomics remains the most advanced of the “omics” technologies (Table 1). Current methodologies allow thousands of polymorphisms to be determined from a single DNA sample, or the level of mRNA for thousands of genes to be quantitated simultaneously using microarray technology. Although the results from such microarrays may be very noisy (the coefficient of variation for the repeated measure of mRNA levels is rarely better than 20%), they contain considerably more information about the current status of the organism than individual classical biochemical methods

<u>“omics”</u>	<u>Variation studied</u>
Genomics	Genetic variation
Transcriptomics	mRNA levels
Proteomics	Protein levels
Immunomics	Antibody repertoire
Metabonomics	Low mol wt metabolites
Glycomics	Polysaccharide structures
Ionomics	Ionic composition

**TABLE 1: A selection of established “omics” and what they measure.**

Despite the current focus on genetic factors as the source of the variability in human disease phenotypes, it is already clear that the pattern of expressed proteins contains has information from the pattern of expressed genes. All kinds

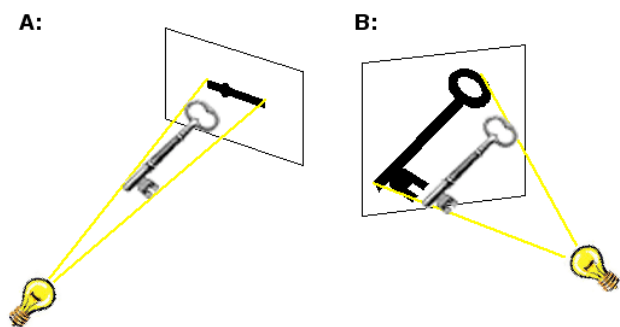


**FIGURE 1: A schematic representation of any multivariate dataset.** For classical multivariate analysis (such as multiple regression,  $N$  must be greater than  $K$ , the  $K$  variables should be noise-free and uncorrelated and the  $X$ -matrix should be complete. For MVA,  $K$  can be much larger than  $N$ , the  $K$  variables can be multicollinear and the  $X$ -matrix noisy and incomplete.

of post-translational modifications result in a proteomic dataset which has some correlation with the genomic dataset from the same individual but which also has additional information to contribute to the descriptor. Much of this difference lies in the different timescales over which gene, mRNA and protein levels vary (DNA sequence only varies within an individual at very specific loci – such as the immunoglobulin genes – and due to rare somatic mutation events, whereas mRNA levels vary on a shorter timescale integrating genetic and environmental inputs, and protein levels may vary slowly or quickly depending on the protein in question – elastin molecules can last a lifetime, while many cell cycle regulatory proteins last only a few minutes).

The range of “omic” technologies now available to construct a descriptor of the individual is increasing all the while. Metabonomics (the study of low-molecular weight, usually non-peptide, metabolites such as sugars and lipids) is rapidly becoming established. Immunomics is a term we have used to describe the study of the repertoire of antibodies expressed by an individual, while ionomics is the term given to the study of the ionic milieu of a cell or tissue. The output from all of these “omics” studies, potentially each applied to several different tissues (such as blood and urine), can be combined to generate a broad descriptor of the individual. Such a descriptor might have tens or even hundreds of thousands of variables within it, reducing to effectively zero the chance that any two individuals will ever have the same descriptor.

While the technology for creating these descriptors has raced ahead, the challenge remains to interpret them in a useful way. There is no practical possibility of overviewing such a large data matrix visually in order to pick out patterns that might be associated with a given phenotype. Instead, automated pattern recognition



**Figure 2: The impact of choosing the right projection.** A projection is a lower dimensional representation of a complex higher dimensional structure. Here, the 2-dimensional projection of a 3-dimensional key is shown. In panel A, a projection is chosen which poorly reflects the variance of the original 3-dimensional object, and the resulting 2-dimensional representation is not even identifiable as a key. In panel B, the optimum projection is chosen which maximises the retention of the information in the original 3-dimensional object in the 2-dimensional representation, which is now clearly recognisable as a key. PCA is a mathematical method to select the optimum projection of the X-matrix into A principle components where  $A \ll K$ .

methods will be required. The problem is that the “omics” dataset violates the requirements for traditional modelling approaches such as multiple regression.

The “omics” dataset is typically “short and wide”, with thousands of variables measured for only a smaller number of individuals, and does not therefore have the full rank required by classical statistics (Figure 1). Worse still, the component variables are often very highly intercorrelated (a problem called multicollinearity). For example, in a typical metabonomics dataset, the average pairwise intercorrelation of the variables is above 0.9. Finally, the “omics” dataset is characteristically noisy and often incomplete with significant missing data.

*Megavariate Analysis (MVA) offers a solution to these problems*

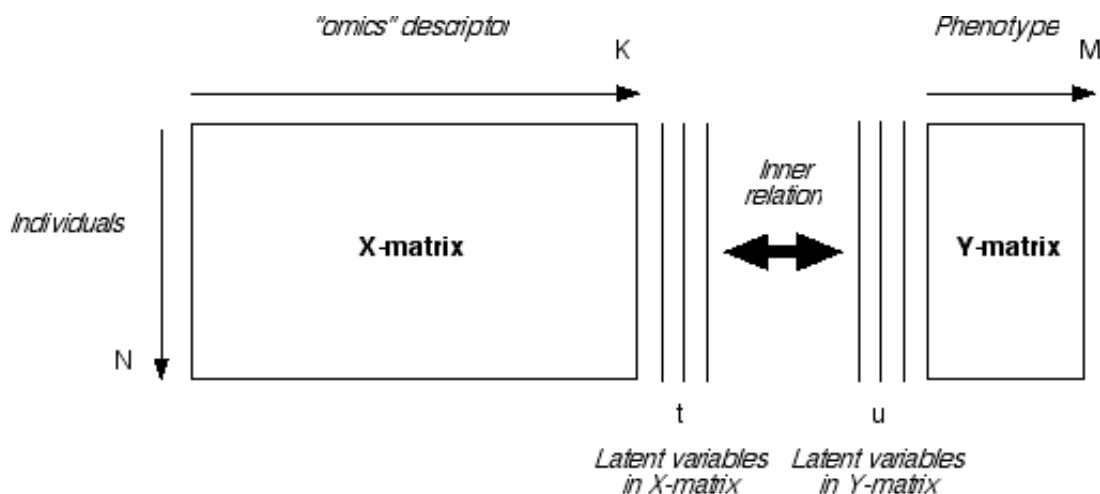
Fortunately, there is a solution to this problem. There exists an alternative to classical statistical treatments that was developed in the early part of this century by Hermann Wold and colleagues at the same time that many classical statistical methods were being developed. However, while classical statisticians often focussed on datasets that impose unrealistic assumptions on the nature of real-world dataset, Megavariate Analysis (MVA) methods were based on modelling real data. Today, many fields have embraced the power of MVA: in chemical engineering it has been used to optimise yields from chemical plants; in medicinal chemistry it has been used for quantitative structure activity relationship (QSAR) studies; in process engineering it has been used to detect faults in

semiconductor fabrication production lines at an early stage. In biology, from epidemiology to medical diagnosis, however, it has largely been overlooked in favour of the traditional approaches. The overwhelming size and complexity of the “omics” descriptor is only now providing the irresistible driving force towards the adoption of MVA methods in biology.

The principle behind the MVA approach lies in approximating a large number of correlated variables with a smaller number of so-called latent variables that together describe most of the variation in the larger dataset. This has the advantage of reducing the complexity of the data and in most cases yields a more robust description of the variation encoded within the dataset since random noise does not contribute to the implicit correlation structure of the variables. The use of latent variables is also relatively unaffected by missing data points.

There are numerous variations on the approach, the simplest of which is Principal Component Analysis (or PCA). PCA provides an overview of a megavariate dataset by extracting one or more principal components (or latent variables) that describe the majority of the variance within the dataset. Clustering of the individuals composing the dataset may now be obvious, and might lead to interesting observations about the phenotypic properties that underlie the clustering. PCA is essentially a projection method: consider a large and complex 3-dimensional object such as a key. We can represent the key by various simplified 2-dimensional representations. Clearly, some 2-dimensional representations represent better approximations of the 3-dimensional key than others. Consider Figure 2a, where the 2-dimensional projection is taken along the long axis of the key shaft. The result is sufficiently uninformative that you would not guess you were looking at a key. In Figure 2b, the projection is changed, and now represents the optimal 2-dimension projection of the key, and the resulting pattern is instantly recognisable. PCA mathematically selects the projection that maximises the description of a K-dimensional matrix (where K is the number of variables in the descriptor) in just a handful of dimensions (usually one, two or three). A detailed description of the PCA approach (as well as other MVA techniques) is beyond the scope of this review, but is discussed in detail by Wold and colleagues [1].

PCA provides an overview of the dataset but it does not relate the phenotype of the individual to the measured parameters. For example, we may want to diagnose the extent of heart disease (measured by coronary angiography) from an “omics” descriptor. The basic design of such an experiment is shown in Figure 3. Now we have an X-matrix composed of K variables making up the “omics” descriptor from N individuals, just as in Figure 1. However, now we also have a set of M measurements

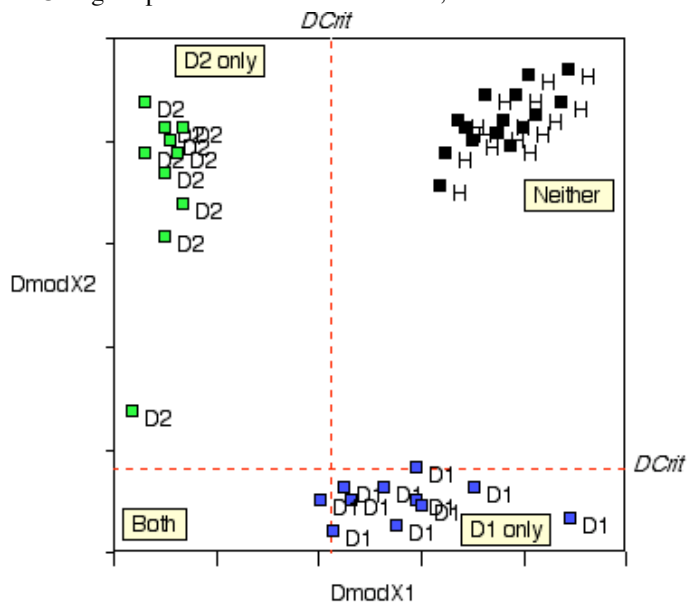


**FIGURE 3:** A schematic representation of a PLS experiment. For a PLS experiment, a Y-matrix is constructed from M variables which describe the phenotype of the individual, in addition to the X-matrix made up of K variables from the "omics" descriptor. PLS performs projections analogous to PCA, describing the X- and the Y-matrix in terms of latent variables, with the added restriction that the latent variables from X (t) should describe the maximum variance in the latent variables from Y (u). This restriction yields the "inner relation" between t and u.

describing the phenotype of the N individuals, making up a Y-matrix. How do we relate the two? MVA provides a technique for experiments such as this, called PLS (projection to latent structures by partial least squares). In principle, PLS performs a PCA analysis on the Y-matrix to yield a small number of latent variables, and then constructs a series of latent variables from the X-matrix which explain the maximum variance in these Y latent variables. PLS models allow us to relate our detailed "omics" descriptor to the real world, and provide the first insight as to which component variables provide the strongest association with a particular phenotype. Such relationships may be the first insight into new biological mechanisms that might reveal important aspects of the pathophysiology of the disease, or act as future therapeutic targets.

For medical diagnosis, however, we need the ability to construct a model that completely separates diseased individuals from healthy ones, and which is subsequently able to predict the health status of an individual from the "omics" descriptor alone. Variations of the MVA techniques described here allow just such predictions to be made: PLS-DA is a variation of PLS in which a dummy Y-matrix is constructed using the class membership data of the known individuals. This model is then used to predict the Y score for individuals of unknown phenotype using just their "omics" descriptor. This works well when the classes are homogeneous and tightly defined, but it may be a demanding assumption for real world classes such as a disease that might result from a range of distinct molecular causes. An alternative approach is called SIMCA (Soft Independent Modelling of Class Analogy). Here, a PCA

model is constructed of each class independently, and each unknown individual is fitted to each of the models in turn. Using a parameter called DmodX, we can assess the



**FIGURE 4:** An example of a Cooman's Plot. Individuals with one of two diseases (D1 or D2) or who are healthy (H) are tested against two PCA models, one for D1 (model X2) and one for D2 (model X1) in a SIMCA procedure. The distance from each model is computed (DmodX1 and DmodX2 respectively) and plotted. The Dcrit value for each model defines the class. Thus, there are four possible outcomes: an individual could be predicted to have D1 only, D2 only, neither or both. In this example, the healthy individuals are correctly predicted to have neither disease D1 or D2.

likelihood that any given individual of unknown phenotype is actually a member of that class. For two classes, that raises four possible outcomes: the individual could be a member of either class, both or neither. The best representation for this is the Cooman's plot (Figure 4), which may represent the output of choice for medical diagnostic tests of the future.

Such prediction methods are potentially tremendously powerful. But there is a danger: it is possible to overfit a model to the “omics” descriptors because of the very large number of variables and the inherent complexity of the dataset. Such a model might explain the class membership of all the known samples very well, yet have no real predictive power for individuals not used to construct the model. The only protection against such overfitting is to perform experiments with a rigorous external validation phase, in which the class membership of entirely new individuals are tested against the models, and the predictive power of the model quantitatively monitored.

*Diagnosing heart disease: the first steps towards a clinically useful diagnostic based on MVA*

In 2002, we published one of the first examples of the application of MVA to an “omics” descriptor for medical diagnostics [2]. Our input dataset was a metabonomic profile obtained by 600MHz proton NMR spectroscopy of human serum samples. The NMR spectrum, containing over 32,000 datapoints, was first data-reduced to about 200 integral bins that represent a crude metabolic profile of the individual. This 200-point “omics” descriptor was collected for about 80 individuals, 40 with severe heart disease diagnosed by coronary angiography, and 40 with healthy coronary arteries. PCA of the data revealed good separation of the two groups, and we subsequently built a PLS-DA model that could predict the disease status of individuals not used to construct to the model from their NMR spectral characteristics alone.

While this study highlighted the enormous potential of MVA-assisted diagnosis based on “omics” descriptors, it is merely the first step on a long road to bringing such tests into the clinic. We built a model comparing heavily diseased individuals with healthy individuals, which is likely to maximise the separation between the classes. We are currently refining the method to apply to the whole population of people suspected of having heart disease, and the outcome of this second, larger study (called MaGiCAD) will determine whether MVA-assisted diagnostics are ready to make the transition to the clinic [3]. Fortunately, our initial study relied on a very crude “omics” descriptor, with just 200 highly collinear variables all derived from a single analytical technique on a single tissue sample. There is plenty of scope for increasing the diagnostic, and even prognostic, capability of the method by increasing the inherent detail of the “omics” descriptor.

*A vision of the future for medical diagnostics*

Where will MVA-assisted diagnosis based on “omics” descriptors take us? Our pilot studies in a range of diseases, including heart disease, hypertension, diabetes, Alzheimer’s disease, osteoporosis and osteoarthritis suggest that the approach has broad applicability irrespective of the disease area. It is not difficult to imagine that within five years it will be possible to automatically generate a multimodal “omics” descriptor with thousands of variables containing genomic, proteomic, immunomic and metabonomic information. Indeed, for some of these techniques the necessary equipment may become small and cheap enough to sit in the office of a primary care physician. The resulting “omics” descriptor would then be fed into an expert computer system with an MVA-based classification capability (perhaps based around SIMCA, in which the descriptor is fitted in turn against models of a wide range of different diseases). The expert system would then provide the physician with a range of probabilities for the presence of each of the diseases. With a sufficiently detailed “omics” descriptor it is possible that most, if not all, of the predictions would be near certain – “0.000” for diseases not present and “1.000” for diseases present.

Such a machine would revolutionise medical provision. For many prevalent diseases (such as coronary heart disease or osteoporosis) preventative treatment is possible providing the onset of the disease can be detected before clinical symptoms are apparent. Akin to the well-established application of MVA to early fault detection on the semiconductor fabrication production lines, a similar approach would provide an early warning of deviation from normality and treatment could be initiated. For metabonomics, Jeremy Nicholson has likened this to a “metabolic trajectory” for individuals through life, which can be monitored, and if necessary manipulated with drugs. The same concept applies to the more general “omics” descriptor – regular monitoring leading to early fault detection and subsequent correction to normality.

Despite all this potential, MVA continues to meet with resistance from classical statisticians uncomfortable with the idea of modelling noisy, multicollinear datasets that are not of full rank. MVA has been slow to gain acceptance at the expense of the orthodox approach, but ultimately the proof of the method lies in its predictive capacity. Once one is able to make clinically useful diagnoses using MVA-assisted analysis of “omics” descriptors, then the validity of the approach is proven, irrespective of the theoretical arguments. In the face of such explicit evidence of its utility, the burden will lie on the guardians of the orthodox approaches to update their theoretical arguments to fit the facts: MVA works and it is here to stay.

*Acknowledgements* We are very grateful to Svante Wold (University of Umea and Umetrics) for his help and support educating us in the theory and practice of MVA, and its application to medical diagnostics. We are also grateful to Professor Jeremy Nicholson, Dr Elaine Holmes and their colleagues (Imperial College of Science, Technology and Medicine) for collaborating with us on the application of metabonomics to medical diagnostics.

#### FURTHER READING

[1] Eriksson L, Johansson E, Kattaneh-Wold N & Wold S. Multi- and Megavariate Data Analysis. *Published by Umetrics Academy, Umea, Sweden.* pp1-527 (2001)

[2] Brindle J, Antii H, Holmes E, Tranter G, Nicholson J, Bethell H, Clarke S, Schofield P, McKilligin E, Mosedale D & Grainger D. Rapid and non-invasive diagnosis of the presence and severity of coronary artery disease using proton-NMR-based metabonomics. *Nature Med.* **8**: 1439-45 (2002).

[3] <http://www.magicad.org.uk> [This website provides additional information about the MaGiCAD study which is currently running to test the clinical utility of MVA-assisted diagnosis of heart disease using a metabonomic descriptor].

---

**IMPORTANT NOTE:** This manuscript was published as part of an occasional series of mini-reviews on the website of the Inflammation Research & Therapy Laboratory at the University of Cambridge (<http://www.graingerlab.org>). The manuscript was therefore not submitted to peer review and only represents the opinion of the author(s). If you wish to cite the information contained in this review, please do so as a personal communication from the author(s) and refer the reader to the web address above. The copyright of this manuscript, including all figures, remains with the author(s). However, they explicitly permit this manuscript to be freely distributed (either electronically or in hardcopy) subject to the sole restriction that it is not modified in any way.