Mega-variate genetics: What you find is what you go looking for

Clive E Bowman
School of Biological Sciences, The University of Reading
Whiteknights, Reading, RG6 6AH, UK
Correspondence to: c.e.bowman@reading.ac.uk

Abstract

The inherent subjectivity or 'purpose-dependency' in the epistemology of discovery from measuring biology is posed using exemplars from high dimensional medical genetic research. The human observer is integrally embedded in the numerical and graphical representation of simple or complex biology via choice of: ascertainment, numerical measure, referential basis, kernel, learning method, feature search and analogous cognitive display of such quantitation. Reality under reason is polythetic. Parsimony constrained reductionism is operational not intrinsic - individuality infers that the *gestalt* is supreme. Despite the fact that 'there is nothing but what you can think there is' - Popper rules!

Key words

aesthetic; complexity; fit; human choice; model; Occam's razor; relativism; simplicity; sufficiency

Presented at: *19th Altenberg Workshop of Theoretical Biology. Measuring Biology - Quantitative methods: Past and Future.* Konrad Lorenz Institute, Austria 12th September 2008. Latex version: Friday, February 6, 2009

Introduction

Science operates by the abstraction, from experience, of commonalities to produce theory or postulates. These condensates are instantiated as formal or informal models which are testable by experiment. Pythagoras said 'Number is the measure of all things' (Hamming 1980) - thus measurement and quantification is integral to the process of understanding biological 'form and function'.

Modern day biology has been seduced by the proposition that 'measuring more is better' (rather than measuring better) and with the explosion of technological instrumentation diverse data on individual organisms is being collected at an exponential rate. High dimensional numerical investigations (modest n - rows or observations, very large p - columns or variables) are everywhere whether in chemistry, biology or sociology (see H Martens and M Martens presentation at this meeting, Eriksson et al. (2006a) and Eriksson

et al. (2006b)). In that way, as Einstein (1944) wrote: *So many people today and even professional scientists seem to me like someone who has seen thousands of trees but has never seen a forest.* Understanding the wood i.e. a *system* of trees (see de Marco, Gordon and Hallgrimsson presentation at this meeting) is the key to biological story-telling. Now, whilst - up to a certain point - knowing the definition of a tree helps if one is trying to spot a wood, an ever increasing clarity about trees does not help - rather having more instances of woods does!

Nevertheless, the concept of needing numerical data to justify conclusions (at least in medicine - Huth (2008)) goes back at least three centuries (Gavarret 1840). Statisticians have long dealt with multivariate data (Krzanowski 2000; McLachlan 1992), the challenge now is this so-called mega-variate data ($p >>> n$, see for example Eriksson et al. (2002); Grainger (2003); Lee et al. (2003); Rubingh et al. (2006); Vis et al. (2007)). Is extracting principled knowledge, of the message that Nature is transmitting, from this current modern-day deluge of atomist data any different from that of the past?

In the field of pharmaceutical genetics a popular version of such data is referred to as 'whole genome scans' (Romero 2002; Roses 2004). Such will be used as illustrative examples of epistemology in this treatise to show that mega-variates are not 'exceptional' (*ėxcĕ′ptional a.* forming an exception; unusual - Sykes (1982)). In these medical exemplars will be used the divergence framework posed by Bowman et al. (2006); Delrieu et al. (2005); Delrieu and Bowman (2006, 2007) - popularising the work of Robin Sibson (Jardine and Sibson 1971). These log likelihood ratios have useful properties including a Bayesian calculus (see Curtis et al. (2007)). They can be manipulated with linear algebra for each observation (i.e. each individual), group etc (Charalambous et al. 2008). The pre-processing arithmetic methodology posed replaces each (exponentially distributed qualitative or quantitative) data point with a marginal entropic 'sufficient' (Fisher 1922) measure on $\Re$ for the question that is being asked which that individualised observation of the variate contributes. The data X is first summarised and non-linearly mapped per axis, as in the Appendix, to a natural non-arbitrary additive information space dependant upon X (and its estimated parameters $\hat{\theta}$) to form a topological space (see Mitteroeker presentation at this meeting). The method deploys these projected measures (which represent the magnitude - not the significance - of the information that each observation contains resolving the uncertainty (Bharath 1987) of the contrast or dialectic of interest) as a point-wise homology function or 'map' for 'data replacements' whilst retaining the original second (and higher) order structure of the biological measurements (i.e. the *pattern* of correlation or covariation in the original space). Euclidean pursuit decomposition of the latter resultant deformed matrices (Huber 1985) to yield 'spectral' signatures (Delrieu and Bowman 2007), as well as the measures themselves (Delrieu et al. 2005), have a direct interpretation in terms of a Popperian (Popper 2002) objective. Being information based, the approach would be an appropriate algebra for the challenge of abduction over the 'landscape' of multiple data types (see Harald Martens in this volume) and joint rhetorics (see Schaefer presentation at this meeting). Operationally this simultaneous geometric approach avoids issues of significance (see McCloskey presentation at this meeting) and repeated testing. In this marriage of Kantian antimonies (see Darvas presentation at this meeting), just as in the Fisherian synthesis (Fisher 1918), the method offers a rationalisation between the dual Mendel-Bateson particulate and Galton-Pearson biometric views of genetics.

Nature is macroscopically quantised - fundamentally instantiated by individual distinct organisms that can be collected, observed and quantified. As in physics (aka Heisenberg uncertainty principle), the act of observation disturbs all biological systems. An isolated system only *represents* biological reality (see de Marco in this volume). Taxonomists use collections to develop models and hypotheses of how the world is organised just as other scientists use their precision equipment for the same purpose (Kohler 2006). As such, ascertainment is crucial in determining what a mega-variate investigation finds out just as much as the *Weltanschauung* ('World-view') the investigator holds. Classes of such individuals are axiomatically only describable in polythetic terms as, even after ignoring measurement error, no two organisms (bar clones in equivalent environments) have the same total set of characteristics. Polytheticsm (Wittgenstein 1953) could be considered as an uncertainty principle for monotheticsm or a stochastic version of it. The individual nature of such biological observations in elevating the importance of ascertainment requires the all-too-often forgotten axioms of random sampling (Kolmogorov 1956). Whilst replication is the (often subconscious) *sine qua non* of mega-variate science (Ioannidis et al. 2001; Lohmueller et al. 2003; Hirschhorn 2002; Vieland 2001), often scant regard is made over what sampling frame this is or under what shared (constrained or unconstrained) convention is 'acceptable replication' (see Galileo's relativity principle in Darvas in this volume).

With the advent of pervasive high-performance computing, Man is entering the age of massive arithmetical manipulation. What can be gained from this? Nothing - if it cannot be imagined (or logically derived - see Darvas presentation at this meeting). Common experience is evidence that what you know (your memory) determines what you notice

(your perception). When you take a swig of coffee only to taste someone-else's tea, it is disgusting - even if you love tea. The expectation (memory) of coffee blended with the reality (sensation) of tea makes the tea taste different - we sense in a relative way. Free-will or stochastic indeterminacy (see de Marco presentation at this meeting) also play a part. Even so-called 'objective' scientific investigations under the Popperian paradigm are a relativistic act - models of reality are transfer functions of input to output (Figure 1) and interpreted within our *Weltanschauung* (see for instance the duality of wave-corpuscle representations in physics - Darvas presentation at this meeting). Their numerical identifiability via plausible reasoning (Jaynes 2003) requires the operation of an abstraction or 'filter' (designed in the context of stochastic measurement for the question of interest - i.e. constructed around observed number and the theory or Popperian postulate for refutation). This filter defines the structure of the meaning in the realised answer to the question (Figure 8).

There is a subjectivity of choice of the measure, the filter and the aggregate as well as a simplicity or approximation (relative truthfulness) of using linear nets to model a fundamentally non-linear reality (see de Marco presentation at this meeting). Numbers and quantification are a human construct (see Nunez presentation at this meeting) and there are many grades of subjectivity (or shared hallucinations!) - see Darvas presentation at this meeting. Apparently simple phenomena such as the double pendulum (see Bowman and Delrieu (2008)), epilepsy and immunological responses hide non-linear mechanisms. The fundamental indeterminism of over-determined small sample mega-variate problems (Charalambous et al. 2008) further requires the subjective choice of regularisers often with recourse to complexity constraints - operationally one has to use Ockham's razor (Ockham 1495). Filters or aggregates themselves require, respectively, relevant referential nulls or referential ontologies - these only pre-exist subjectively. Such conceptual tools are culture dependent (see Nunez this volume). The choice of the null is crucial in determining the confidence in the numerical outcome of any model of biology. Ontologies ('functional shells' - see Oxnard presentation at this meeting) are in some sense conditional smoothers. Predictions are always probabilistic in biology (see de Marco presentation at this meeting).

Megavariate insight is gained not just by the manipulation of number and reason but recognition via displays and graphs - codified quantitative decompositions or condensates of simple rules or geometries - with such attributes as: position, direction, distance, trajectory, orthogonality, additivity, inclusion, node components, edge relations, edge

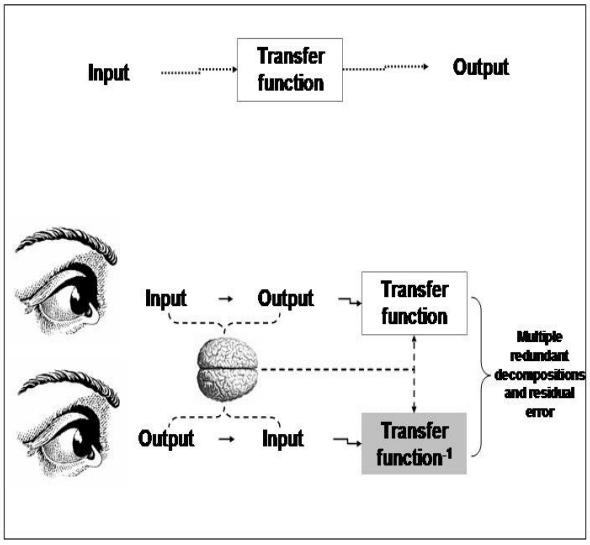


Figure 1. The art of science is in designing an observational experiment to restrict the non-identifiability of the hidden or latent transfer function (*TF*) decomposition. 'Objective' tools for that art are Ockham's Razor (Ockham 1495) and the Popperian paradigm (Popper 2002) but are subjectively dependant upon the investigator's *Weltanschauung* and are subject to Kuhnian change (Kuhn 1996). Dotted lines indicate abduction. Input and output can be physical or notional flows. *TF* may be a physical cybernetic-informational process or an abstract functional fractal system (see de Marco presentation at this meeting). *Upper*:- Epistemological concept; *Lower*:- The projection of input space into output space (aka deduction, or prediction) mapping of transfer function; versus, The projection of output space into input space (aka induction or inference) mapping of inverse transfer function (TF^{-1}).

weights, noise characteristics etc. One could consider numbers to equal each 'tree' in a close-up view - with multivariate condensed displays being the distant 'forest'. However, directions of effects depend upon ethnography (see Nunez this volume). We assume that we see the world just as it is, but we are constrained by what we experience (and hold in memory) - comprehension ultimately comes from heuristics and analogy.

Ascertainment

Megavariate genetic studies are notorious for their lack of replication (Ioannidis 2007). This failure of researchers to repeat and generalise genetic-association findings is most commonly attributed to:- poor genetic coverage, measurement error (Ioannidis 2007), insufficient statistical power (Bacanu et al. 2000), poor false discovery control (Wakefield 2007), population stratification (Hinds et al. 2004) bias, or various forms of between-study heterogeneity (Chen and Witte 2007; Skol et al. 2006) or environmental influences

(Lasky-Su et al. 2008) - see Figure 2. Galileic regularities (aka 'replication' - see Darvas presentation at this meeting) is repeated agreement under deliberate independent experimental challenge (of place, time and system of reference) as to it's falseness. Only a rigorous understanding of ascertainment will rescue investigators from this bane of poor replication in mega-variate genetic linkage and association studies.

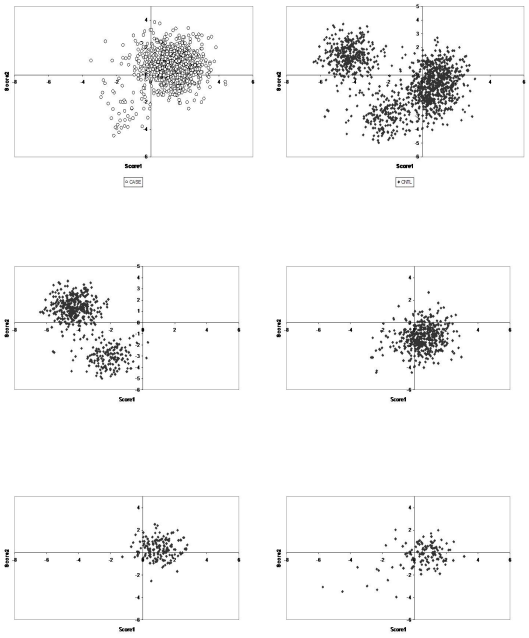


Figure 2. Megavariate ordination example of metabolic disease phenotype >7000 SNPs (tri-state single nucleotide polymorphisms), >1500 genes over the whole genome (Roses et al. 2005), eigen decomposition of correlations of observed divergences (see Delrieu and Bowman (2006)). Score plot - individuals closer on plot are genetically closer. Horizontal axis represents case-control distinction. Open diamonds = cases. Closed diamonds = control individuals. Note genetic distinction aliased with ascertainment. *Top left*: Plot showing homogeneous cases; *Top right*: Plot showing control genetic heterogeneity; *Middle Left*: Controls from investigational centre 1; *Middle Right*: Controls from investigational centre 2; *Bottom Left*: Controls from investigational centre 3; *Bottom Right*: Controls from investigational centre 4.

In pharmaceutical drug target discovery and clinical validation both genetic linkage and genetic association studies have been used widely (Roses 2004). Practice with both these types of study has frequently been out-dated and flawed. Objective *independent* repeatability is the Rosetta stone of modern scientific belief. However, experimental replication i.e. the repeated occurrence of a genetic result (whether found by linkage or association) in other independent samples - which is the desired outcome to foster confidence that the results are real - rarely occurs in practice using these techniques (see for linkage - Baron (2001) or for association - Cardon and Bell (2001)). This results in false hopes and wasted medical resources.

A common denominator in both experimental strategies is ascertainment - the selection and procurement of the mega-variate genetic sampling unit, albeit whether this unit is a diseased family (c.f. linkage pedigree studies) or a diseased patient (c.f. clinical trial association studies). Neglecting careful consideration of ascertainment is an often-made mistake - which will ensure the continued failure of the replication of results of potential medical importance.

Current genetic linkage studies use *specially* chosen sets of families of willing volunteers selected by interested investigators. In this way they are no different than a museum zoologist using pedigrees of particular butterflies *deliberately* collected in-flight for showing an interesting spot (c.f. 'picking winners'). The experimental units (families) are neither comprehensive nor representative of the diversity of humans with and without the disease of interest spread worldwide - just as flying spotty butterflies are not a good sample of global lepidopterans. This matters - if one wants a high chance of replicating a linkage finding (see *informal* argument below).

Similarly, current clinical trial mega-variate genetic association studies invariably use (specially selected and restricted) *ad hoc* 'opportunity' populations (Gonik and Smith 1933) of people who, whilst being collected carefully by committed physicians, do not represent or cover diseased and non-diseased humans worldwide (just like sets of different coloured stamps collected by a hobbyist philatelist from letters that they receive are not necessarily comprehensive either). This matters - if one wants a high chance of replicating an association finding (see *informal* argument below).

Ascertainment in mega-variate science, both in genetic and non-genetic studies (see Senn (1997)), is currently a practice often modeled upon the Victorian era - a time of well-meaning gentlemen collectors and amateur hobbyists - it needs to be regularly deployed as a modern-day rigorous science.

Ascertainment can, of itself, be a fixed choice or a random observable in an experiment but the lessons concerning replication are the same (see *informal* argument below). This argument herein uses just two ascertainment sets for clarity of exposition (it can be generalised for n sets: $(S_1 \ldots . S_n)$. The notation below uses A for association and L for linkage but either letter (or the mapping to the English words) could be used interchangeably.

Ascertainment - as a fixed choice

Consider this *informal* thesis:-

Imagine in an experiment e, one observes the conditional probability $p(A|S_e)$ as high, where p indicates probability,

| means 'given', A is association (or linkage) and S_e is the controlled choice of ascertainment for that experiment.

This is a 'positive result' that one hopes to independently replicate. In doing so, what one actually really wishes to conclude is that the unconditional p(A) i.e. probability of positive association (or linkage) *irrespective* of the ascertainment (that is, the *a priori* choice of the particular S), is high.

Now axiomatically by *extending the conversation*

$$p(A) = p(A|S_e).p(S_e) + p(A|\bar{S}_e).p(\bar{S}_e) \quad (1)$$

where $\bar{S}_e$ means 'not that ascertainment'; and, S_e and $\bar{S}_e$ are mutually exclusive and exhaustive ascertainments that independently cover the whole space of S.

In modern-day science we aspire to use the practical concept of 'experimental replication' of association (or linkage) i.e. $p(A|S_i)$ is large for all chosen i , to infer that the unconditional p(A) is high.

Now, $p(A|S_n) = \frac{p(S_n|A).p(A)}{p(S_n)}$ by Bayes Theorem (for n = {e, 'not e'}), and thus from (1)

$$p(A) = \left[\frac{p(S_e|A).p(A)}{p(S_e)}.p(S_e) + \frac{p(\bar{S}_e|A).p(A)}{p(\bar{S}_e)}.p(\bar{S}_e)\right]$$

which on canceling and rearranging trivially yields

$$p(A) = p(S_e|A).p(A) + p(\bar{S}_e|A)p(A) \quad (2)$$

where the terms $p(S_e|A) + p(\bar{S}_e|A)$ must sum to 1.

Now, one can improperly *ensure* - for fixed p(A) - that the first compound term in (2) is high if the 'probability' of the chosen ascertainment given association is high (c.f. 'picking *a priori* winners'). This then is equivalent in (1), (given no overall preference for any particular ascertainment i.e. $p(S_e) = p(\bar{S}_e)$ - that is the ascertainment is typical and representative of the whole of S), to $p(A|S_e)$ being high i.e. a 'self fulfilling prophecy' of positive experimental association (or linkage). A deliberately 'biased' sample *will* ensure success! Looking at it another way, if '$p(S_e)$' is unusually high i.e. S_e is not typical nor representative of the whole of S, then - for any fixed $p(A|S_e)$ - the first compound term in (1) will be high leading to an incorrect conclusion regarding the probability of experimental association (or linkage) if someone inadvertently assumes that random ascertainment had occurred. A biased sample *will* ensure erroneous conclusions!

Either way, if the first compound term in (1) or (2) is improperly high (by having 'picked winners' or equivalently having an unrepresentative experimental ascertainment), then in order to maintain the axiomatic equality, the second compound term in (2) - for fixed p(A) - will be low *per force* as there must be a lower probability of the alternative ascertainment given association (or linkage) occurring. In this case, equivalently, the second compound term in (1) must be low. This can only be by either the probability of the other ascertainment being low i.e. the alternative ascertainment is also not typical nor representative of the whole of S or $p(A|\bar{S}_e)$ is low i.e. the converse result fails to independently replicate in practice!

The informal arguments in the preceding paragraphs follow in the same way if one starts from a position of low values for the first compound terms rather than high.

Confidence of replication can thus be best achieved,

- If '$p(S_i|A)$' is similar for all i, i.e. no ascertainment is chosen *a priori* by conscious or unconscious design over any other when the association is occurring - an 'unbiased choice',

and corrolaritively,

- If all ascertainments are (equally) likely through random sample selection comprehensively over the whole sampling frame - $p(S_i)$ is similar (over S) for all i.

Ascertainment - as a random observable

Consider this *informal* thesis:-

Imagine in an experiment e, one observes the joint probability $p(L\&S_e)$ as high, where p indicates probability, & means 'and', L is linkage (or association) and S_e is the *observed* random ascertainment occurring for that experiment.

This is a 'positive result' that one hopes to independently replicate. In doing so, what one actually really wishes to conclude is that the marginal p(L) i.e. probability of positive linkage (or association) *irrespective* of the ascertainment (that is, the particular *a posteriori* S_e occurring), is high. This is again to be inferred by practical experimental replication of the linkage (or association) i.e. $p(L|S_i)$ is large for all i.

Now axiomatically

$$p(L\&S_e) = p(L|S_e).p(S_e) = p(S_e|L).p(L)$$

and thus by Bayes Theorem

$$p(L) = \frac{p(L|S_e).p(S_e)}{p(S_e|L)}$$

Also, axiomatically by *extending the conversation*

$$p(L) = p(L|S_e).p(S_e) + p(L|\bar{S}_e).p(\bar{S}_e)$$

So,

$$p(L|S_e).p(S_e) + p(L|\bar{S}_e).p(\bar{S}_e) = \frac{p(L|S_e).p(S_e)}{p(S_e|L)}$$

or, by rearrangement

$$p(L|\bar{S}_e) = \frac{p(L|S_e).p(S_e)}{p(\bar{S}_e)}.\left[\frac{1}{p(S_e|L)} - 1\right] \quad (3)$$

Now, for a fixed p(L) and a given large $p(L|S_e)$, the probability of experimental linkage (or association) given the alternative ascertainment ($p(L|\bar{S}_e)$) will only be large i.e. replication may occur to some degree, when,

the multiplicative term in (3) above i.e. $\frac{p(S_e)}{p(\bar{S}_e)} \cdot \left[\frac{1-p(S_e|L)}{p(S_e|L)}\right]$ (4)

is not small.

Now, by definition,

$$1 - p(S_e|L) = p(\bar{S}_e|L)$$

as S_e and $\bar{S}_e$ cover the whole space S.

Inserting this into the square bracket in (4), shows that

- Independent experimental replication may occur to a degree if $\frac{odds(S_e)}{odds(S_e|L)}$ is not small.

This odds ratio is intuitively sensible and is best achieved, for fixed p(L) and any uncontrolled $p(S_e)$, by ensuring that odds(S_e|L) = odds(S_e), that is the relative probability of ascertainments given linkage (or association) is the same as any overall relative probability of ascertainments. This is when,

- The experimental ascertainment is *unbiased* with respect to linkage (or association) being present (i.e. 'winners' are *not* picked, odds(S_e|L) = 1 for all e),

and

- When no particular ascertainment is preferred over any other (i.e. random sampling over whole sampling frame, odds(S_e) = 1 for all e),

as then $\frac{odds(S_e)}{odds(S_e|L)}$ = 1 and $p(L|S_e) = p(L|\bar{S}_e)$ for all e ($\equiv$ full replication since S_e and 'not S_e' are mutually exclusive and exhaustive of the whole ascertainment space S).

Ascertainment - sampling

Irrespective of whether the ascertainment (i.e. the selection and procurement of experimental units) is an *a priori* choice or an *a posteriori* observable, the *informal* argument above shows that :-

(1) Megavariate genetic linkage studies, where the selection of that pedigree (amongst all pedigrees) is *because* inheritance by descent has been (partly) seen (c.f. 'Butterfly collecting'), will be beset with a lack of success in experimental replication due to potential biased ascertainment.

(2) Clinical trial mega-variate genetic association studies, using opportunity samples (c.f. 'Stamp collecting') - no matter how carefully collected and executed - will similarly fail to replicate experimentally if they lack appropriate coverage of the whole sampling frame of relevance.

Accordingly, the best strategy to follow in order to mitigate the potential lack of replication of mega-variate genetic results of potential importance is,

- *The unbiased choice of experimental units randomly over the whole potential sampling frame that the medical conclusion is to be made over.*

Megavariate science needs to leave the 'ethic of Victorian collectors' behind it. In practice, this means a (more) random choice of pedigrees from the whole set of possible human pedigrees in genetic linkage studies (i.e. treat S_e as random). It also means appropriately randomly drawn samples from epidemiological studies or surveys rather than using *ad hoc* clinical trial collections for mega-variate genetic association studies (i.e. a more representative S_e). Some institutes are leading the way with the latter - Academia Sinica in Taiwan use geographically stratified, population density weighted samples for their control genetic epidemiological collection in association studies (Pan et al. 2004), the Wellcome Trust Case-Control Consortium drew epidemiological samples UK-wide (Wellcome Trust Case Control Consortium 2007) etc. Only then will the bane of the lack of replication be slayed.

This is nothing new (see proof in Kolmogorov (1956)) just often-forgotten even in the well-trodden fields of medicine (see the work of Cochrane - Chalmers (2008)) and meta-analysis (Senn 1997). Does it matter? - yes it does. Delrieu and Bowman (2006) gives an example where the mega-variate genetic distinction between disease cases and controls is clearly related to the method of ascertainment - subjects being genetically distinct simply because they were referred to contributing medical centres rather than volunteering.

Ascertainment - choice of measure

The choice of measure (aka summary data statistic) is just as crucial and the also oft-forgotten meaning of $E[log(likelihood)]$ versus $log(likelihood)$ makes the link here. The reasoning behind the use of observed log likelihood versus expected log likelihood is invariably presented solely in a statistical manner in scientific treatises. However, log likelihood is clearly intended for *this* sample here (irrespective of it's typicality) whilst the expected log likelihood is clearly intended for a *typical* sample of which this *is* one.

For instance the curvature with respect to ML parameters $\hat{\theta}$ estimated in a mega-variate genetic model as

$$[var(\hat{\theta})]^{-1} = -\frac{\delta^2 log(likelihood)}{\delta\theta^2}|_{\hat{\theta}}$$

depends upon the patterns of *these* individuals one has and is susceptible to ascertainment bias whether that is

- from the sampling (aka collection) process (say, an unplanned observational *ad hoc* imbalanced with uncontrolled confounders approach, versus, a randomised planned balanced and orthogonally designed-to-control confounders approach), or,
- from the intrinsic biological basis itself (say, inbred versus outbred humans; population structured samples versus not structured etc).

The curvature with respect to ML parameters estimated in a mega-variate genetic model as

$$[var(\hat{\theta})]^{-1} = -E[\frac{\delta^2 log(likelihood)}{\delta \theta^2}]|_{\hat{\theta}}$$

is the *typical* value, that is *only* appropriate if this *was* a random unbiased sample of it's 'parent' group. The stochastic expectation operator E is acting itself as an *aggregator* to move one from the individuals' level to that of the group. This (long-run property) is only valid under the assumption of no confounders (i.e. random sampling).

In that way, divergences analysis implicitly assuming sample $log(likelihood) \neq E[log(likelihood)]$ is only sensible if there *are* other factors. Such will find mega-variate patterns that can be traced back to the ascertainment process. This is nicely shown by the opportunity sample in the example in Figure 3 - the difference between the loci inferred by decomposing $log(likelihood)$ ratios versus those inferred form using $E[log(likelihood)]$ ratios illustrates the hidden ascertainment factors. What you find is what you go looking for.

Similarly, divergences analysis implicitly assuming sample $log(likelihood) \equiv E[log(likelihood)]$ is only sensible *if* one can guarantee there are *no* other factors. This is nicely exemplified by the work of Voight and Pritchard (2005) where cryptic relatedness had usually a negligible impact upon the number of so-called 'false positives' (driven by over-dispersion) in *well-run* genetic association studies of outbred populations i.e. *despite* the biological bias, good sampling stochastics allow $log(likelihood) \cong E[log(likelihood)]$.

Inappropriate sampling biases the features found - ascertainment matters. Repeatability needs randomness. Investigators should beware of any hidden random sampling assumption. But what is the sampling frame for such random sampling?

Filtering

Referential basis

Supervised (and semi-supervised) learning require a reference ontology. This is true of divergences analysis (Delrieu et al. 2005; Delrieu and Bowman 2006) as much as for phylogenetic cladistics (Hennig 1966) - conclusions can only be made with reference to an out-group (cf. 'controls'). Shared primitive characters (symplesiomorphies) have no intrinsic power to determine phylogeny - only the sharing of (derived) apomorphies. These are determined by human judgement. In case-control mega-variate genetics - it is 'case-ness' (the dissymetry from controls - see Figure 15) that is measured and decomposed i.e. the degree of distinction from the reference not the agreement with it. In that way one, is measuring the degree of 'falseness' to a posed 'null' - the evidence which that observation falsifies attribution to the reference population.

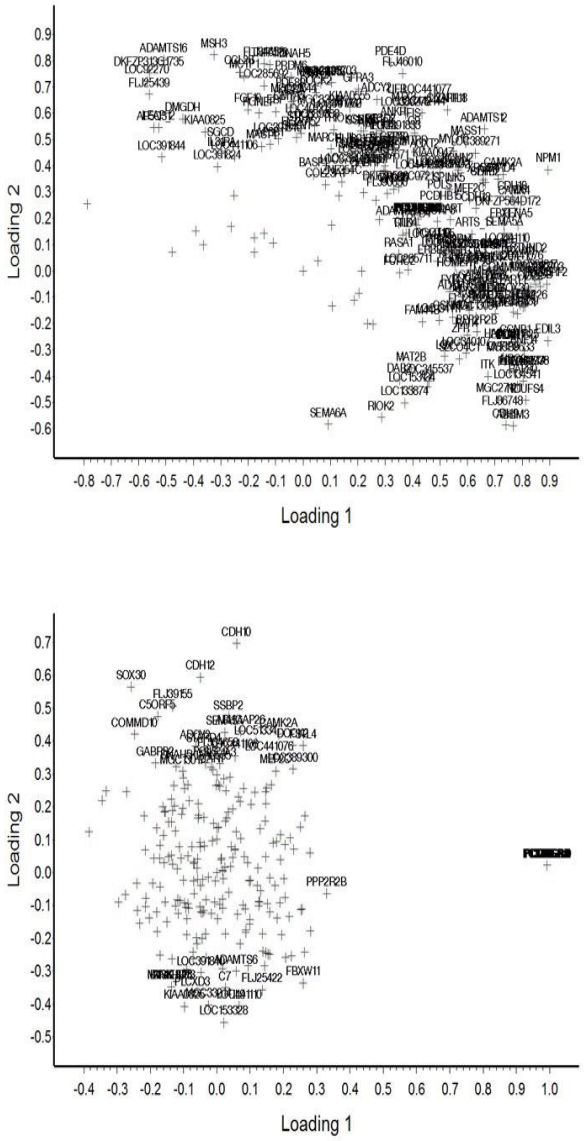


Figure 3. Megavariate ordination example of chromosome 5, Affymetrix 25000 SNP chip, phenotype = % weight loss on investigational drug treatment. Gene loadings plot, correlation matrix, horizontal axis is treatment-placebo direction. If this sample was typical (lower plot) - only one locus (protocadherin) on the far right from thousands is important for this trait - but in fact (upper plot) very many loci show the ascertainment signal. *Upper*: Observed log likelihood ratio $0.5 * (\ln(\sigma_2^2) + \frac{(y-\mu_2)^2}{\sigma_2^2} - \ln(\sigma_1^2) - \frac{(y-\mu_1)^2}{\sigma_1^2})$ - see Bowman et al. (2006); Delrieu and Bowman (2006), correlation SVD triage. This sample's inferred physiological network components are legion as both between and pooled within population components are found. *Lower*: Expected log likelihood ratio $0.5 * (\ln(\sigma_1^2) + \frac{\sigma_2^2}{\sigma_1^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_1^2} - \ln(\sigma_2^2) - 1)$ - see Bowman et al. (2006), correlation SVD triage. The typical or expected inferred physiological network components if this was a random representative sample are few as only between population components are found.

Figure 2 nicely shows that depending upon which centre's subjects would be taken as representative of controls, different genetic distinctions would be found. The reference group matters in mega-variate science. In fact, what (real or non-real) reference group should be used?

- One group - reference group $\hat{\theta}_r$?
- Pooled over every group's data - reference group $\hat{\theta}_.$?
- Group average over each group's parameters - reference group $\bar{\hat{\theta}}$ (i.e. treat as information radii)?

all are subjective choices according to purpose.

What is the null for divergences analysis? A permutation empirical null of no location difference (i.e. $\theta_i = \theta_z$ as in http://taxonomy.delrieu.org)? But these network nodes are triaged by single value decomposition (SVD), so what should be the null for covariance or correlation matrices (or kurtosis etc for ICA - see Koch and Naito (2007))? A uniform matrix (or tensor)? A random walk based upon historical initial conditions (see Bookstein presentation at this meeting)? A fully-mixed random field? A column-wise shuffle? A bootstrapped sample from the data? These are all subjective choices depending upon purpose (see Appendix).

The issue is not what is the evidence *for* this? Rather where is any evidence *against* this result? It is the discrepancies that matter - finding a conceptual projection that systematically pulls these out from the 'noise' - so the null matters. Imaginative criticism is required - (Popper 2002) and an art in neglecting the irrelevant circumstances or options (see Darvas and Bookstein respectively presentation at this meeting). Taking the simplest null of 'no assumptions' may be easy but may be incorrect. Context matters, for instance, as Wardrop (2008) recounts - searching for a unifying single cause of disease in the elderly for a given set of symptoms would be pointless; as it is statistically more probable, especially in an ageing patient, that multiple yet independent disease processes occur to account for an unusual set of symptoms as opposed to a single "rare as hen's teeth" diagnosis. Unthinking application of Ockham's razor is wrong: Ockham $\neq$ (just) simple (see account of the helicocentric versus geocentric view in Darvas presentation at this meeting). Symmetry may not appear until the right basis frame is posed (see account of Maxwell in Darvas in this volume).

So, what is the sampling frame? Figure 4 shows that ignoring the historical basis of biological individuals can lead to a confounded background to any apparent random sampling. traditional statistics relies upon exchangeability (Jaynes 2003). Similarly in network analysis (and covariance matrices can be mapped to networks via walk laplacians - (Chung 1994)), estimation techniques are even more complicated as the structure of a random sample seldom

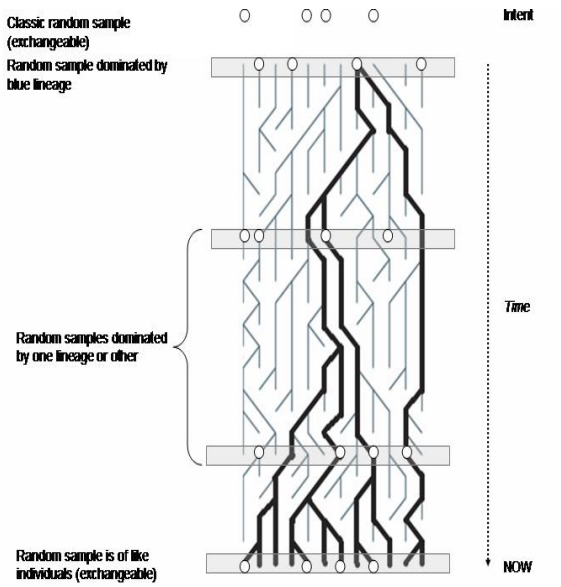


Figure 4. Coalescent tree for individuals showing genealogy into the past versus intent of a simple random sample. Random samples (grey slices with white circles) are not of exchangeable (Jaynes 2003) individuals unless evolutionary time is taken into account. The current population is not representative of the total set but descends from a single ancestor (lineage in bold). Figure adapted and enhanced from Bamshad and Wooding (2003).

matches that of the overall network (see Figure 5). Boundary specification (Wasserman and Faust 1994) may seriously affect the structure of a discovered network - the domain (sampling frame); the context; and thus, ascertainment matters.

This is nothing new, Bookstein presentation at this meeting points to the success of statistics in the exploration of **a**historic effects. However biology quintessially is path dependent (i.e. contingently frozen) - the 'story' from the past matters in understanding current and future form and function.

Kernel choice

Co-occurrence (aka *pattern*) matters - evolution works upon suites of characters i.e. upon the covariance space not individual characters in isolation (Bowman and Delrieu 2008). Humans particularise the space into individual characters. Each individual evolutionary unit is a diplotype (two haplotypes) in evolutionary space. This diplotype represents the carriage of physiological components or network nodes and their mutual interactions (antagonism, agonism, synergism, anti-synergism, control, release etc). The number and size of nodes drive penetrance (via 'density') and pleiotropy (via 'scattering') under environmental inputs ('torch light' - see Figure 6). Measuring biology is an attempt to deconvolve

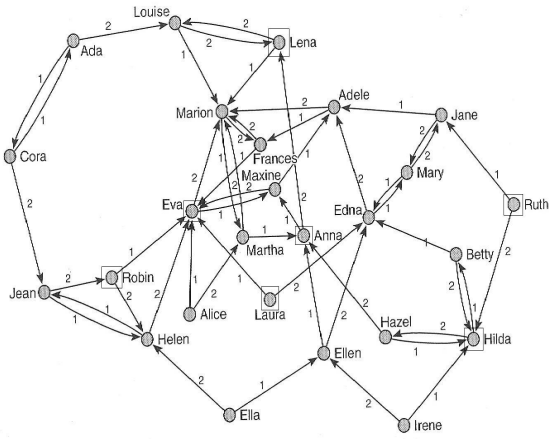


Figure 5. Illustrative network of first (1) and second (2) preferences for, say, a kidney transplant amongst a set of women. A random sample will not estimate the variability of edges if nodes (structural components) are missing - consider a network sample of Anna, Eva, Hilda, Laura, Lena, Robin and Ruth highlighted by squares. One finds fewer choices per person than the two choices in the overall network for the simple reason that choices to women outside the sample are neglected. For such problems of boundary specification and sampling also see Wasserman and Faust (1994). Adapted from a sociogram of dining-table partners in de Nooy et al. (2005).

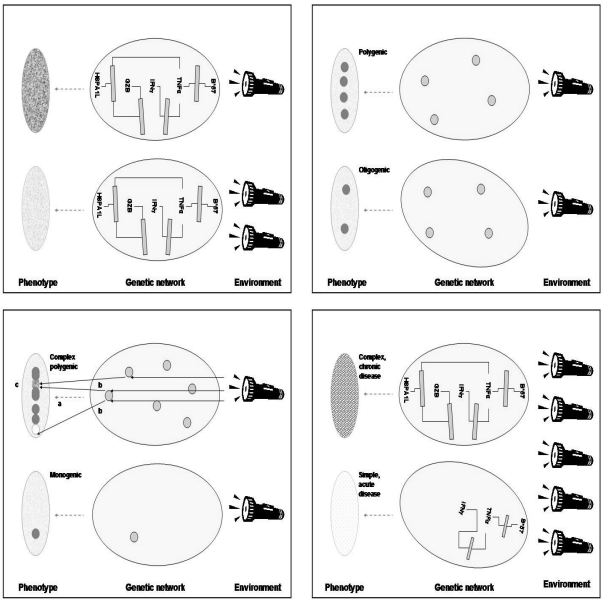


Figure 6. Genetic network density, node size and orientation in determining complexity of apparent phenotype, genotype penetrance and pleiotropy (see also Bowman and Delrieu (2008)). Grey bars = epistatic interactions in underlying physiological network. *Top Left*: Analogy of density of shadow depending upon strength of torchlight - Differing environmental strengths between individuals could yield differing penetrance for the same genetic network basis; *Top Right*: Analogy of the complexity of phenotype depending upon the overall number (impact) of shadows - Depending upon genetic orientation, a trait can appear polygenic or oligogenic as loci move out and in of linkage disequilibrium (LD); *Bottom Left*: (a) Analogy of shadows of balls under torchlight - Differing complexity in a genetic network could yield different phenotypes for the same environment. (b) Analogy of multiple internal reflections - Each genetic determinant has pleiotropic effects. 'Hub' nodes should show greater pleiotropy. (c) Analogy of interference patterns of internal reflections - Each genetic determinant may combine with those of others. An infinite number of such balls randomly arranged could engender a normally distributed trait; *Bottom Right*: Polythetic trait - The complexity of the phenotype depends upon the variability of environmental input, the size and orientation of the genetic network, the impact of network nodes and epistasis. The *gestalt* matters (Bowman and Delrieu 2008).

such relations from observation in the context of the question being asked whether simple or complex (see Gordon and Hallgrimsson in this volume) despite inherent indeterminancy (see de Marco in this volume).

Any filter is a distorting subjectivity, since if a phenomenon is truly mega-variate then trying to characterise it with a single number (or a few numbers) may result in a loss of information if a non-sufficient statistic is used. There is rarely only one reasonable way of combining multiple quantitative measurements into a single number (Greenspan 2007). Aggregating multiple values into one requires weighting the values corresponding to different dimensions and such weightings almost always involve subjective or at least 'purpose-dependent' choices. Moreover, there is a subjectivity of feature space according to the choice of kernels in a filter as illustrated by different divergences - see Figure 7.

Learning methods such as maximum variance (=PCA, SVD), maximum kurtosis (Skillicorn 2007) etc are pre-defined feature search algorithms. They triage variates by a model 'fit' to an objective function. Much has been written regarding SVD and principal components although some recent results (Hoyle and Rattray 2004, 2007; Patterson et al. 2006) have not been widely promulgated. Independent components analysis and combinations of dimensional reductions methods for mega-variate data have been proposed (Koch and Naito 2007). The literature on these topics is vast - eigen analysis having an illustrious history in genetics with Cavalli-Sforza (see Patterson et al. (2006)). Preprocessing using wavelets (Daugman 2003) or other linear and non-linear aggregates as a simplifying convenience is often employed. These are in some sense regularisers or smoothers (Charalambous et al. 2008).

Megavariate observation of biology, as any finite measurement, is imprecise - in a Platonic world it makes 'mistakes'. One can only give an approximation to reality by measuring biology. Under multiple redundancy (Figure 1):

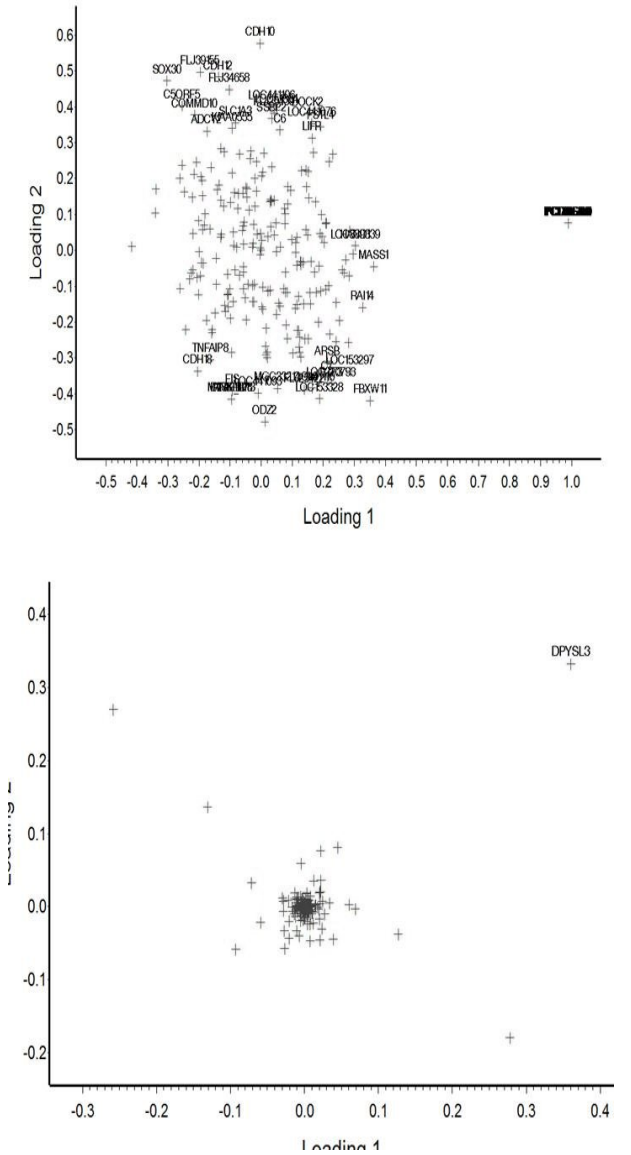


Figure 7. Megavariate example of chromosome 5, Affymetrix 25000 SNP chip, phenotype=% weight loss on investigational drug treatment. Gene loadings ordination, horizontal axis is case-control direction. Upper: JSDL - Expected over groups of expected log likelihood ratio, $0.25(*\frac{\sigma_1^4+\sigma_2^4+(\sigma_1^2+\sigma_2^2)(\mu_1-\mu_2)^2}{\sigma_1^2\sigma_2^2}-2)$ decomposition - see Bowman et al. (2006), correlation matrix SVD triage - see Delrieu and Bowman (2007). This shows commonality with Figure 3 *Lower* but also shows distinctions as the divergence is now not *directed*). Lower: Expected log likelihood ratio $0.5(*\ln(\sigma_1^2)+\frac{\sigma_2^2}{\sigma_1^2}+\frac{(\mu_1-\mu_2)^2}{\sigma_1^2}-\ln(\sigma_2^2)-1)$ - see Bowman et al. (2006), covariance matrix SVD triage, showing how variate scale in the kernel affects the results (compare to Figure 3 *Lower*) - only DYPSL3 (from the Ulip family of phosphoproteins) is found to dominate the decomposition - see Delrieu and Bowman (2006).

Numquam ponenda est pluralitas sine necessitate - 'Plurality ought never be posed without necessity' Ockham (1495) - that is, one should accept the explanation with the fewest number of *assumptions*. But, as Exupéry (1939) states: *Il semble que la perfection soit atteinte non quand il n'y a plus rien à ajouter, mais quand il n'y a plus rien à retrancher* ('Perfection is not when there is nothing to add, but when there is nothing to remove'). Thus the best explanation is not necessarily the simplest - it is the aesthetic of fit (aka elegance or symmetry, plus 'fit-for-purpose') not the lack of complexity *per se* that matters, see Wardrop (2008). As Einstein (1933) said: *It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.* In practice, a competent investigator will arrive at a fairly good unifying approximation whilst incompetent ones will be far off the mark (see Bookstein presentation at this meeting). As with medicine (Joosse and Pormann 2008), however, the fault does not lie with mega-variate biology but with:-

- the particular circumstances which may be too complex to be fully encompassed by (simple) general principles (see the Saint's Triad in Wardrop (2008)), and/or,
- the practitioners themselves, who as individual human beings, make mistakes.

This fundamental duality in residual heterogeneity (after a model of biological reality has been fitted) of whether it is:-

- 'lack of fit' to the chosen model, versus
- stochastic noise

requires a *deus ex machina* subjective symmetry breaker. Either faith in an invariance (see Darvas presentation at this meeting) or an abstract theory or the application of an assertion about error noise properties is needed. For the latter, either the logical context of the phenomenon (e.g. the defence to a lie that she had been under sniper fire in Bosnia in the assertion by Hilary Clinton during the 2008 Democrat nomination in the USA that it was not surprising that she got some things wrong, seeing how she spoke millions of words a day - is conceptual bunkum!); or a chosen axiomatic context (e.g. sphericity Delrieu and Bowman (2006)); that determines 'signal' from 'noise'. Without it, one enters into irresolvable debates such as the long-standing arguments over what should be the metric for between-study heterogeneity in the meta-analyses of randomised clinical trials (Engels et al. 2000). Subjectivity and human choice are at the nub of resolving the inexact nature of biology.

This is in contrast to the use of statistical distributions to *notionally* categorise individuals into classes. Whilst use of a sufficient (Fisher 1922) statistic encapsulates the essence

of a grouping (up to the point of stochastic variation), individuals are intrinsically unique - they are not 'errors' or 'mistakes' of an 'ideal'. The construct of an average person, is just that - an artificial construct - as shown by common experience with clothes sizes and fit - such a typical person does not exist (see Bookstein presentation at this meeting). The *gestalt* is unique and immensely indeterminate (see discussion on Gaddis and Elsasser by Bookstein in this volume).

Just as there is no such thing as a 'pure significance test' (that is without a specified alternative) there is no context-free biological question. Enquiry has a context. In science (and government) this must be under the pre-eminence of the Popperian framework (Popper 2002). In Gulf War 2, it is clear that the policy was to attack Iraq and evidence that Saddam Hussein had weapons of mass destruction capable of hitting the West within 45 minutes was exhaustively searched for to justify it, rather than governmental policy being made in the light of the contrary evidence found. Despite the subjective relativism of human endeavour, theories (or postulates) are only falsifiable, not provable. Although choosing the hypothesis or a mathematics that would - if true - best explain the relevant evidence (Lipton 2004) is a sensible idea, just because a hypothesis or a mathematics is reasonable (or personal - see Darvas presentation at this meeting) is not sufficient to make it accurate or right (Zinkernagel 2007). More credit should be given to a theory or a mathematics when data are *predicted* rather than accommodated, and the best explanations are those that 'most increase our understanding' (Lipton 2004). Yet the algebra of progress must remain Popperian - experimental test and discrepancy (dissymetry) is the King!

Cognition

Numbers are numbers and simply have order and scale. Pattern must be *recognised* or actively sensed (see M Martens in this volume). Humans recognise 'lumps and bumps' - condensates - versus a uniform null not versus random fluctuations. In fact, human cognition is poor at detecting randomness (see Bookstein in this volume) - often seeing inhomogeneities in such. The cognitive null is the uniform not stochastic randomness - feature recognition in mega-variate data by humans thus depends upon structured display of 'fluctuations' versus 'stasis', not structure versus randomness. The surprising properties of random walks (see Bookstein presentation at this meeting) are poorly comprehended.

Eigen decomposition (SVD) is a 'least-squares' procedure of the data itself (aka the observed *design* of data variables) and produces an orthogonal decomposition which by maximising variance will graphically show ellipsoid groupings if present. Is nature comprised of (a few) orthogonal latent vectors? Certainly such representations can be mapped via a walk laplacian (Chung 1994) to a (simple) inferred network relationship between items (see Pirmohamed et al. (2007)) that has seductive cognitive properties. Moreover, scientists are inculcated with plots. Nevertheless both representations are limited by their linearity, additivity, superposition etc. Operationally they are *useful* arithmetical reductionist instruments for numbers but they do not necessarily reflect the *actual* structure of biology. Human insight relies upon analogy (see Figure 6) as 'there is nothing but what you can think there is'.

Practical limitations force the reductionist use of linear models (Ioannidis 2007) - sample sizes for adequate power to detect interactions are often prohibitively large (Hunter 2005). Accordingly, evidence on identified and large-scale replicated gene-gene and gene-environment interactions for instance is limited (Garcia-Closas et al. 2005). Interactions themselves are context specific (Zhong and Sternberg 2006) - lack of interaction in a multiplicative model is an interaction in an additive model. Metric matters. Evolution works upon suites of characters i.e. upon the covariance space not individual characters in isolation. Moreover it is the *interaction* of such characters - 'to give more bang for the buck' over and above their carriage on the haplotype that is important in defining the physiological interaction of biological components.

Imagination, creative invention, serendipity (Meyers 2008), 'attention' (in a Buddhist sense Pirsig (1999), 'flow' - in a psychological sense Csikszentmihalyi (1996)) and curiosity, as much as reasoned intellect are the watchword in mega-variates because one cannot solve problems by using the same kind of thinking as was used when they were created. Above all, mega-variate medicine is polythetic (see Pirmohamed et al. (2007) for example). In practice, Man is a geometer - measuring biology with ever-increasing numbers of numbers, yet even with objective reason - mega-variate science is still human purpose-dependent.

Wrap-up

Whether in simple or complex systems (see de Marco, Gordon and Hallgrimsson presentation at this meeting), mega-variate genetics is not exceptional. Megavariate genetics is simply evolutionary demography. The inherent contingent subjectivity and purpose-dependency in this story-telling from measuring biology can be summarised in Figure 8 - 'What you can find is what you can imagine you should go looking for'. Where it came from and how you go looking for it (or *rather* looking for its converse or postulated disagreement - i.e. imaginative criticism Popper (2002)) determines what

you find. What you accept as found is what is beautiful and useful within your *Weltanschauung*.

In that way measuring form and function in biology remains an interplay of Platonic intellect and Aristotelian operations. The world which we see and in which we live is derived from:-

- the world of ideas (Plato Plato (428-347BC)) or stable laws;

which we explore by proceeding from:-

- unstable individual observable instances - the particular or nominal;

to the general (Aristotle Aristotle (384-322BC)).

There is science in the art of mega-variates but an art in the science of mega-variates.

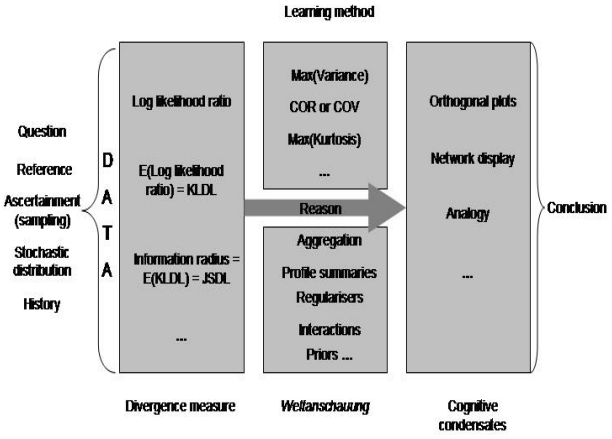


Figure 8. Numerical epistemology for biological story-telling - How meaning can be extracted from observation through divergences. Human choices in the subjective filtering of mega-variate reality via latent vectors to yield insight - 'there is nothing but what you can think there is'. The basis of knowledge of hidden causes is:- data, experience, reason and analogy as in medical epistemology (Joosse and Pormann 2008). Data can be ordinal or cardinal. The question is the objective. Ascertainment is the systematic collection to answer that posed contrast of interest over the sampling frame or domain of relevance. Model regularisers (Charalambous et al. 2008) can be elastic or plastic, phenomena or epiphenomena (see Bookstein presentation at this meeting). They can be: complexity penalties, inter-point distances, initial conditions (origin), time paths (temporal history), space trajectories (spatial history), or 'functional shells' (see Oxnard this volume). Maximum variance (second moment) learning leads to ellipsoid fuzzy groups. Maximum kurtosis (fourth moment) learning leads to non-ellipsoid discrete widely-spaced groups - see also Delrieu and Bowman (2007).

Acknowledgements

I thank my wife and daughter for their tolerance and endless patience with me writing often at strange hours and locations. Alun D McCarthy for asking me in 2004 why it is striking that disease genetic linkage studies often fail to replicate. The anonymised examples arise from work at GlaxoSmithKline R&D under Allen Roses. Without my close friend (and one-time colleague) Olivier Delrieu, MD - this treatise would never have seen the light of day outside of my turbulent mind.

References

Akaike H (1974) A new look at the statistical model identification *IEEE Transactions on Automatic Control* AC-19, 6, 716-723

Aristotle (384-322BC) http://en.wikipedia.org/wiki/Aristotle (28 August 2008)

Bacanu S-A, Devlin B and Roeder K (2000) The Power of Genomic Control *Am. J. Hum. Genet.* 66:19331944

Bamshad M and Wooding S P (2003) Signatures of natural selection in the human genome *Nature Reviews Genetics* 4, 99-111

Baron, M (2001) The search for complex disease genes: fault by linkage or fault by association? *Molecular Psychiatry* 6, 143-149

Bharath R (1987) Information theory *Byte* 291-298

Bowman C and Delrieu O (2008) Immunogenetics of drug-induced skin blistering disorders *Pharmacogenetics* (submitted)

Bowman C, Delrieu O and Roger J (2006) Filtering pharmacogenetic signals In: S Barber, P D Baxter, K V Mardia and R E Walls (*Eds.*) *Interdisciplinary Statistics and Bioinformatics*. University of Leeds, 41-47

Cardon, L R and Bell, J I (2001) Association study designs for complex diseases *Nature Reviews Genetics* 2, 91-99

Chalmers I (2008) Archie Cochrane (1909-1988) *J R Soc Med* 101: 41-44

Charalambous C, Delrieu O and Bowman C (2008) Whole genome scan algebra and smoothing In: Barber, S, Baxter P D, Gusnanto, A and Mardia, K V (eds) *The Art and Science of Statistical Bioinformatics* Univ. of Leeds 21-27

Chen G K and Witte J S (2007) Enriching the Analysis of Genomewide Association Studies with Hierarchical Modeling *The American Journal of Human Genetics* 81:397404

Chung F R K (1994) Spectral Graph Theory *Regional conference series in mathematics* CBMS Conference on Recent Advances in Spectral Graph Theory held at California State University at Fresno, June 6-10, 1994

Csíkszentmihályi, M (1996) *Creativity: Flow and the Pyschology of Discovery and Invention* Harper Perennial, New York

Curtis D, Vine A E and Knight J (2007) A pragmatic suggestion for dealing with results for candidate genes obtained from genome wide association studies *BMC Genetics* 8:20

Daugman, J (2003) Demodulation by complex-valued wavelets for stochastic pattern recognition *International Journal of Wavelets, Multiresolution and Information Processing* 1, 1-17

Delrieu, O and Bowman, C E (2005) Visualisation of gene and pathway determinants of disease In: S Barber, P D Baxter, K V, Mardia and R E Walls (*Eds.*), *Quantitative Biology, Shape Analysis, and Wavelets*. University of Leeds, 180pp. 21-24

Delrieu, O and Bowman, C E (2006) Visualising gene determinants of disease in drug discovery *Pharmacogenomics* 7 (3) 311-329

Delrieu O and Bowman C (2007) On using the correlations of divergences In: Barber, S, Baxter, P D and Mardia, K V (*Eds.*) *Systems Biology and Statistical Bioinformatics*. University of Leeds pp 27-35

de Nooy W, Mrvar A and Batagelj V (2005) *Exploratory Social Network Analysis with Pajek* Cambridge University Press

Einstein A (1933) *On the Method of Theoretical Physics* The Herbert Spencer Lecture, delivered at Oxford (10 June 1933); also published in *Philosophy of Science* 1, 2 (April 1934): 163-169

Einstein A (1944) Letter to Robert A. Thorton, Physics Professor at University of Puerto Rico (7 December 1944) [EA-674, Einstein Archive, Hebrew University, Jerusalem]. Thorton had written to Einstein on persuading colleagues of the importance of philosophy of science to scientists (empiricists) and science.

Engels E A, Schmid C H, Terrin N, Olkin I and Lau J (2000) Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses *Stat Med* 19, 17071728

Eriksson L, Johansson E, Lindgren F, Sjstrm M and Wold S (2002) Megavariate analysis of hierarchical QSAR data *Journal of Computer-Aided Molecular Design* 16 (10):711-726

Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikstrom C and Wold S (2006a) *Multi- and Megavariate Data Analysis. Part I. Basic Principles and Applications* Umetrics 425pp

Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikstrom C and Wold S (2006b) *Multi- and Megavariate Data Analysis. Part II. Advanced Applications and Method Extensions* Umetrics 307pp

Exupéry A de S (1939) *Terre de Hommes* [translated as *Wind, Sand and Stars*] Chapter III: L'Avion, p.60

Fisher R (1918) The correlation between relatives on the supposition of mendelian inheritance *Philosophical Transactions of the Royal Society of Edinburgh* 52, 399-433

Fisher R (1922) On the Mathematical Foundations Of Theoretical Statistics *Phil. Trans. R. Soc. Lond.* A 222: 309-368

Fogel P, Young S S, Hawkins D M and Ledirac N (2007) Inferential, robust non-negative matrix factorization analysis of microarray data *Bioinformatics* 23, 44-49.

Garcia-Closas M, Malats N, Silverman D, Dosemeci M, Kogevinas M, Hein D W, Tardon A, Serra C, Carrato A, Garcia-Closas R, Lloreta J, Castano-Vinyals G, Yeager M, Welch R, Chanock S, Chatterjee N, Wacholder S, Samanic C, Tora M, Fernandez F, Real F X and Rothman N (2005) NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: Results from the Spanish Bladder Cancer Study and meta-analyses *Lancet* 366, 649659

Gavarret J (1840) *Principes Généraux de Statistique Médicale ou Développement des Règles Qui Doivent Présider à Son Emploi* (Ed.) Bechet Jeune et Labé, Paris 61pp

Gonik, L and Smith, W (1993) *The Cartoon Guide to Statistics* Harper Resource, New York

Grainger, D J (2003) Megavariate Statistics meets High Data-density Analytical Methods: The Future of Medical Diagnostics? *IRTL Reviews* 1:1-6

Greenspan N S (2007) Conceptualizing immune responsiveness *Nature Immunology* 8, 1, 5-7

Hamming R W (1980) The unreasonable effectiveness of mathematics *The American Mathematical Monthly* 87, 2 (February)

Hennig W (1966) *Phylogenetic Systematics* University of Illinois Press [translated by Davis D and Zangerl R]

Hinds D A, Stokowski R P, Patil N, Konvicka K, Kershenobich D, Cox D R and Ballinger D G (2004) Matching Strategies for Genetic Association Studies in Structured Populations *Am. J. Hum. Genet.* 74:317325

Hirschhorn J N, Lohmueller K, Byrne E and Hirschhorn K (2002) A comprehensive review of genetic association studies *Genet. Med.* 4, 2, 45-61

Hoyle D C and Rattray M (2004) Principal-component-analysis eignevalue spectra from data with symmetry-breaking structure *Physical review* E 69, 026124:1-13

Hoyle D C and Rattray M (2007) Statistical mechanics of learning multiple orthogonal signals: Asymptotic theory and fluctuation effects *Physical review* E 75, 016101:1-13

Huber P J (1985) Projection pursuit *Ann. Statist.* 13, 2, 435-475

Hunter D J (2005) Gene-environment interactions in human diseases *Nat Rev Genet* 6, 287298

Huth E (2008) Jules Gavarret's *Principes Généraux de Statistique Médicale*. *J R Soc Med* 101:205-212

Ioannidis J P A (2007) Non-Replication and Inconsistency in the Genome-Wide Association Setting *Hum Hered* 64:203213

Ioannidis J P, Ntzani E E, Trikalinos T A and Contopoulos-Ioannidis D G (2001) Replication validity of genetic association studies *Nat. Genetics* 29, 3, 306-309

Jain A, Kingsley C and Olshen A (2001) An analytical framework and data system for analyzing megavariate biological data in cancer *Nature Genetics* 27:62

Kohler, R E (2006) *All Creatures: naturalists, collectors and biodiversity, 1850-1950* Princeton University Press

Jardine, N and Sibson, R (1971) *Mathematical Taxonomy* John Wiley

Jaynes E T (2003) *Probability Theory: The Logic of Science* [Ed. by Bretthorst G L from 1995 manuscript] Cambridge University Press .

Joosse N P and Pormann P E (2008) Archery, mathematics , and conceptulaizing inaccuracies in medicine in 13th century Iraq and Syria *J R Soc Med* 101: 425-427

Koch I and Naito K (2007) Dimension selection for feature selection and dimension reduction with principal and idependent component analysis *Neural Compuation* 19: 513-545

Kolmogorov, A (1956) *Foundations of the theory of probability* Chelsea Publishing Compnay, New York, 2nd Edition, [translation by Nathan Morrison of *Grundbegriffe der Wahrscheinlichkeitrechnung* in Ergebnisse Der Mathematik (1933)]

Krzanowski, W (2000) *Principles of Multivariate Analysis: A User's Perspective* Oxford University Press 608pp

Kuhn T S (1996) *The structure of scientific revolutions* University of Chicago Press 3rd Edition.

Lasky-Su J, Lyon H N, Emilsson V, Heid I M, Molony C, Raby B A, Lazarus R, Klanderman B, Soto-Quiros M E, Avila L, Silverman E K, Thorleifsson G, Thorsteinsdottir U, Kronenberg F, Vollmert C, Illig T, Fox C S, Levy D, Laird N, Ding X, McQueen M B, Butler J, Ardlie K, Papoutsakis C, Dedoussis G, O'Donnell C J, Wichmann H-E, Celedn J C, Schadt E, Hirschhorn J, Weiss S T, Stefansson K and Lange C (2008) On the Replication of Genetic Associations: Timing Can Be Everything! *Amercian Journal of Human Genetics* in press doi:10.1016/j.ajhg.2008.01.018

Lee K R, Lin X, Park D C and Eslava S (2003) Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method *Proteomics*3(9):1680-6

de Leon A R and Carrière K C (2003) A Generalized Mahalanobis Distance for Mixed Data *Elsevier Science* 1-21

Lipton P (2004) *Inference to the Best Explanation* 2nd ed. Routledge

Lohmueller K E, Pearce C L, Pike M, Lander E S and Hirschhorn J N (2003) Meta-analysis of genetic association studies supports acontribution of common variants to susceptibility to common disease *Nat. Genet.* 33, 2, 177-182

McLachlan, G J (1992) *Discriminant Analysis and Statistical Pattern Recognition* John Wiley & Sons, 544pp

Meyers M A (2008) *Happy accidents - serendipity in modern medical breakthroughs* Arcade Publishing, New York

Ockham W (1495) *Quaestiones et decisiones in quattuor libros Sententiarum Petri Lombardi* (ed. Lugd.),i, dist.27, qu.2, K

Pan, W H, Fann, C S J, Wu, J Y, Hung, S I, Hung, Y T, Chen, Y J, Hsu, C L, Liao, C J, and Chen Y T (2004) Establishment of Taiwan Han Chinese Cell and Gene Bank: Comparing SNP profiles in MHC region with Caucasians. *Presented at the American Society of Human Genetics 54th Annual meeting*. Oct 26-30, 2004. Toronto, Ontario, Canada. (page number: 221 abstract number:1149)

Patterson N, Price A L and Reich D (2006) Population Structure and Eigenanalysis *PLoS Genet* 2(12): e190. doi:10.1371/journal.pgen.0020190

Pirmohamed, M, Arbuckle, J, Bowman, C, Brunner, M, Burns, D, Delrieu, O, Dix, L, Twomey, J and Stern, R (2007) Investigation into the multi-dimensional genetic basis of drug-induced Stevens-Johnson syndrome and toxic epidermal necrolysis *Pharmacogenomics* 8 (12):1661-1691

Pirsig R M (1999) *Zen and the Art of Motorcycle Maintenance* Vintage

Plato (428-347BC) http://en.wikipedia.org/wiki/Plato (28 August 2008)

Popper K (2002) *The logic of scientific discovery* Routledge [translation of *Logik der Forschung* (1934)]

Romero R, Kuivaniemi H, Tromp G and Olson J M (2002) The design, execution, and interpretation of genetic association studies to decipher complex diseases *Am J Obstet Gynecol* 187, 5 : 1299-1312

Roses, A D (2004) Pharmacogenetics and Drug Development: The path to safer and more effective drugs *Nature Reviews Genetics* 5, (9):643-655

Roses A D, Burns D K, Chissoe S, Middleton L and St Jean P (2005) Disease-specific target selection: a critical first step down the right road *Drug Discov Today* 10, (3):177-89

Rubingh C M, Bijlsma S, Derks E P P A, Bobeldijk I, Verheij E R, Kochhar S and Smilde A K(2006) Assessing the performance of statistical validation tools for megavariate metabolomics data *Metabolomics* 2 (2):53-61

Senn, S (1997) *Statistical Issues in Drug Development* John Wiley & Sons, Chichester 442pp

Shannon C E (1948) A Mathematical Theory of Communication *Bell Syst. Techn. J.* 27, Part I: 379-423

Skillicorn D (2007) *Understanding complex datasets. data mining with matrix decompositions* Chapman & Hall/CRC

Skol A D, Scott L J, Abecasis G R and Boehnke M (2006) Joint analysis is more efficeint than replication-based analysis for two-stage genome-wide association studies *Nature Genetics* doi:10.1038/ng1706

Sykes, J B [Ed] (1982) *The Concise Oxford Dictionary of Current English* 7th edition Oxford

Tobler W R (1970) A computer movie simulating urban growth in the Detroit region. *Econom. Geography Suppl.* 46, 234-240

Tribus M (1961) *Thermodynamics and Thermostatics: An Introduction to Energy, Information and States of Matter, with Engineering Applications* D. Van Nostrand Company Inc., New York

Vieland V J (2001) The replication requirement *Nat. Genet.* 29, 3, 244-245

Vis D J, Westerhuis J A, Smilde A K and van der Greef J (2007) Statistical validation of megavariate effects in ASCA *BMC Bioinformatics* 8:322

Voight B F and Pritchard J K (2005) Confounding from Cryptic Relatedness in Case-Control Association Studies. *PLoS Genetics* 1(3): e32 doi:10.1371/journal.pgen.0010032

Wakefield J (2007) A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies *Am. J. Hum. Genet.* 81:208227

Wardrop D (2008) Ockham's Razor: sharpen or re-sheathe? *J R Soc Med* 101: 50-51

Wasserman S and Faust K (1994) *Social Network Analysis: Methods and Applications* Cambridge University Press

Wellcome Trust Case Control Consortium (2007) Geneome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls *Nature* 447, 661-678

Wittgenstein, L (1953) *Philosophical Investigations* Macmillan, New York

Zhong W and Sternberg P W (2006) Genome-wide prediction of C. elegans genetic interactions *Science* 311, 14811484

Zinkernagel R (2007) On observing and analyzing disease versus signals *Nature Immunology* 8, 1, 8-10

Appendix

Concept of divergences

".... everything is related to everything else, but near things are more related than distant things." Tobler (1970)

Let the collection of random variables X over a total of k individuals (the *Data*) be partitioned (according to a potential contrast of interest with respect to a *reference* population z) between r populations $= \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_r \end{bmatrix}, z \in r,$

That is,

$$X_1 = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots x_{1,p} \\ x_{2,1} & x_{2,2} & \dots x_{2,p} \\ \dots & \dots & \dots \dots \\ x_{n,1} & x_{n,2} & \dots x_{n,p} \end{bmatrix} \$$$

(population 1, n subjects, p random variables) and

$$X_2 = \begin{bmatrix} x_{n+1,1} & x_{n+1,2} & \dots x_{n+1,p} \\ x_{n+2,1} & x_{n+2,2} & \dots x_{n+2,p} \\ \dots & \dots & \dots \dots \\ x_{n+m,1} & x_{n+m,2} & \dots x_{n+m,p} \end{bmatrix} \$$$

(population 2, m subjects, p random variables) etc to population r and X_r

Then, given,

$X \sim p[f(\mathbf{X})\Theta]\ d\mathbf{X}$ where $\Theta = \begin{bmatrix} \theta \\ \psi \end{bmatrix}$, and, ψ are nuisance parameters, f is some function with Θ (which is similar over all r populations in our formulation), p is a *joint* probability density and θ are $q \geq r.p$ parameters of interest (for our formulation i.e. each combination of population and variable must have at least one unique parameter). f and Θ define the *model* for stochastic X.

The formulae below are for continuous random variables; for discrete variates simply replace the $\int_{\mathbf{X}_{i,j}} \dots\ d\mathbf{X}_{i,j}$ with $\Sigma_{\mathbf{X}_{i,j}}$

Divergences as natural measures for analysis flow from considering any scientific question to be a contrast. For instance, taking any two populations (i and z) pair-wise:-

The *marginal* self-information (Shannon 1948) for the ith population and jth variable is:

$$-log_2(p[f(\mathbf{X}_{i,j})\Theta_{i,j}]\ d\mathbf{X}_{i,j})$$

This density is a measure of the information content associated with the outcome of a random variable. The unit of this information is the 'bit'. It has also been called surprisal (Tribus 1961), as it represents the "surprise" of seeing the outcome (e.g. a highly probable outcome is not surprising).

The *marginal* differential entropy (or expected surprisal) for the ith population and jth variable for continuous functions is:

$$-\int_{\mathbf{X}_{i,j}} p[f(\mathbf{X}_{i,j})\Theta_{i,j}].log_2(p[f(\mathbf{X}_{i,j})\Theta_{i,j}])\ d\mathbf{X}_{i,j}$$

This information entropy value is a measure of the uncertainty associated with a random variable (it is a continuous analogue of the discrete Shannon entropy). It is a measure in bits of the average information content the recipient is missing when they do not know the value of the random variable.

The *marginal* log probability ratio between the ith population and the reference population z for the jth variable is:

$$log_2(\frac{p[f(\mathbf{X}_{i,j})\Theta_{i,j}]}{p[f(\mathbf{X}_{i,j})\Theta_{z,j}]}\ d\mathbf{X}_{i,j})$$

and is a dimensionless density.

The *marginal* relative entropy (or Kullback-Leibler divergence) between the ith population and the reference population z for the jth variable is a measure (in bits) of the *directed* (or asymmetric) difference between the two probability densities for continuous functions:

$$-\int_{\mathbf{X}_{i,j}} p[f(\mathbf{X}_{i,j})\Theta_{i,j}].log_2(\frac{p[f(\mathbf{X}_{i,j})\Theta_{i,j}]}{p[f(\mathbf{X}_{i,j})\Theta_{z,j}]})\ d\mathbf{X}_{i,j}$$

This value is minus the expected (under population i) log probability ratio. It is the 'loss' when the $p[f(\mathbf{X}_{i,j})\Theta_{z,j}]\ d\mathbf{X}_{i,j}$ density is used to approximate the $p[f(\mathbf{X}_{i,j})\Theta_{i,j}]\ d\mathbf{X}_{i,j}$ density.

The *marginal* squared Jeffrey's symmetric divergence between the ith population and the reference population z for the jth variable is the average Kullback-Leibler divergence over the two groups for continuous functions:

$$-\frac{1}{2}.\left(\int_{\mathbf{X}_{i\&z,j}} p[f(\mathbf{X}_{i,j})\Theta_{i,j}].log_2(\frac{p[f(\mathbf{X}_{i,j})\Theta_{i,j}]}{p[f(\mathbf{X}_{i,j})\Theta_{z,j}]})\ d\mathbf{X}_{i,j}\right) -$$

$$\frac{1}{2}.\left(\int_{\mathbf{X}_{z\&i,j}} p[f(\mathbf{X}_{z,j})\Theta_{z,j}].log_2(\frac{p[f(\mathbf{X}_{z,j})\Theta_{z,j}]}{p[f(\mathbf{X}_{z,j})\Theta_{i,j}]})\ d\mathbf{X}_{z,j}\right)$$

This value is an *undirected* (i.e. symmetric) 'distance' and measures (in bits) the difference between the two probability densities. It is the average 'loss' when each of the densities $p[f(\mathbf{X}_{z,j})\Theta_{z,j}]\ d\mathbf{X}_{z,j}$ and $p[f(\mathbf{X}_{i,j})\Theta_{i,j}]\ d\mathbf{X}_{i,j}$ is used to approximate each other. It is a metric distance if square rooted.

These integrals may have closed antiderivative analytical solutions or can be evaluated by numerical quadrature. Discontinuous densities can be handled by simple conversion to Lebesgue integrals.

Marginal likelihood formulation

Define $L_{i,j}$ as a function, the *likelihood* of the data,

$$L_{i,j}[f(X_{i,j})\Theta_{i,j}]$$

as the *height* of the marginal density $p[f(\mathbf{X}_{i,j})\Theta_{i,j}]\ d\mathbf{X}_{i,j}$ at values over $X_{i,j}$ for variable j and population i.

Then ignoring the sign (by considering it to simply indicate the *orientation* of the manifold $\mathbf{X}_{i,j}$) and replacing the continuous integrals with an infinite summation of Riemann vertical 'slabs' of uniform infinitesimal width (see Figure 9 Upper):-

The homologue of expected surprisal numerically for population i and variable j is:

$$\Sigma_{X_{i,j}} L_{i,j}.log_2(L_{i,j})$$

The log likelihood ratio for population i versus a reference population z for variable j is the *directed*:

$$LLR^z_{i,j} = \Sigma_{X_{i,j}} log_2(\frac{L_{i,j}}{L_{z,j}})$$

It concerns *these* sets *i* and *z* here.

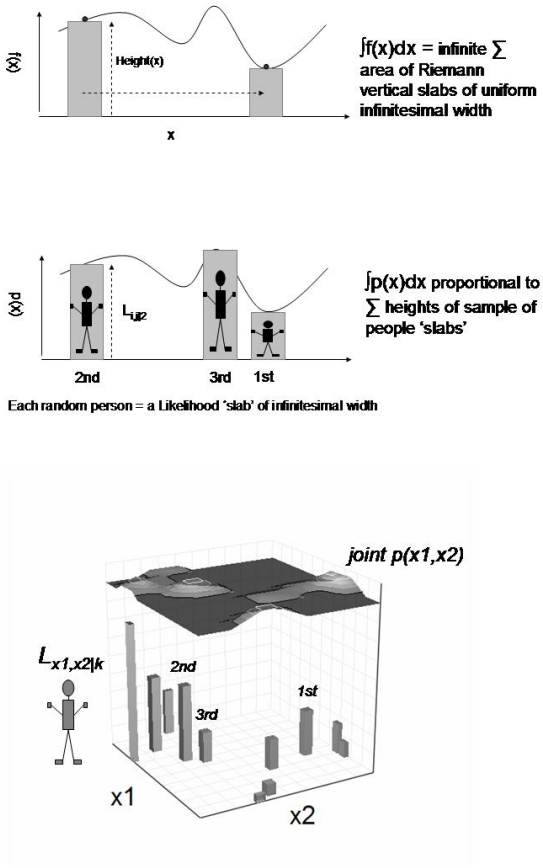


Figure 9. *Upper*: Classic theory of integration of a function f(x). *Middle*: Monte Carlo approximation of a probability density integral. The more people the better the resolution of the function. *Lower*:The individualised likelihoods of a random sample of people can approximate (up to a proportionality constant) a joint bivariate probability density (with correlation structure). Each bar is a subject. Likelihood height exaggerated. Density with contours and shading for clarity. Note: examples here assume perfect knowledge of functional form and parameter estimates.

The expected log likelihood ratio (homologue of the Kullback-Leibler divergence) versus a reference population z for variable j is the *directed*:

$$KLDL^z_{i,j} = \Sigma_{X_{i,j}} L_{i,j}.log_2(\frac{L_{i,j}}{L_{z,j}})$$

As before, it is a homologue of the 'loss' measure when $p[f(\mathbf{X}_{i,j})\Theta_{z,j}]\ d\mathbf{X}_{i,j}$ is used to approximate $p[f(X_{i,j})\Theta_{i,j}]\ d\mathbf{X}_{i,j}$ evaluated each at its parameter estimates. It concerns the *typicality* of *these* sets *i* and *z* here.

The average (over groups) expected log likelihood ratio (homologue of squared Jeffrey's symmetric divergence) for variable j between population i and reference population z is the *undirected*:

$$JSDL^z_{i,j} = \Sigma_{X_{i\&z,j}} \frac{KLDL^z_{i,j} + KLDL^i_{z,j}}{2}$$

It is a homologue of the measure of the average 'loss' when each of $p[f(\mathbf{X}_{z,j})\Theta_{z,j}]\ d\mathbf{X}_{z,j}$ and $p[f(\mathbf{X}_{i,j})\Theta_{i,j}]\ d\mathbf{X}_{i,j}$ is used to approximate each other evaluated each at its parameter estimates. It concerns *these* typical i and z sets *amongst* the sets (i and z) here.

For $p[f(\mathbf{X}_{i,j})\Theta_{z,j}]\ d\mathbf{X}_{i,j}$ and $p[f(\mathbf{X}_{i,j})\Theta_{i,j}]\ d\mathbf{X}_{i,j}$ of the same number of parameters no 'penalty' for change in complexity is needed in these divergences (Akaike 1974). Simple closed forms are available for divergences if data belongs to the exponential family of distributions (Delrieu et al. (2005),Bowman et al. (2006),Delrieu and Bowman (2007)) or mixtures thereof.

In practice, the summation over the $X_{i,j}$ is made via a Monte Carlo approximation (using population parameter estimates) assuming that the individuals are a random sample (see Figure 9 Middle) i.e. for a total sample of k individuals, each measure is a Σ_k of terms containing $\hat{L_{i,j}}|_k$ and $\hat{L_{z,j}}|_k$ (see Figure 9 Lower and Section below).

Parameter estimation and Individualisation

$\hat{\theta}$ are estimated including nuisance $\hat{\psi}$ using maximum likelihood (least squares) or via Bayes theorem using simple conjugate priors. Delrieu and Bowman (2006) uses independent estimation but joint estimation over variates is possible. Estimates can be theory-free treating the observations as a phenomenological 'heap of data' or contingent upon prevaviling mechanistic theories of the underlying phenomena. Conditional characters (Jardine and Sibson 1971) are dealt with by nested dummy variables.

Given parameter estimates for $\hat{\theta_{i,j}}$, $\hat{\theta_{z,j}}$ and $\hat{\psi}$, then it is possible to evaluate any of the likelihood measures at these estimates (i.e. $\hat{L_{i,j}}$) for any individual k (see Figures 10 and 11). For instance numerically calculating the expected log likelihood ratio estimate

$$KL\hat{D}L^{z}_{i,j}|_k = \hat{L_{i,j}}|_k.log_2(\frac{\hat{L_{i,j}}|_k}{\hat{L_{z,j}}|_k})$$

by plugging in the population estimates, together with the *observed* $X_{i,j}$ and $X_{z,j}$ for the kth individual. Leading then, for example, to the estimates

$$KL\hat{D}L^{z}_{i,j} = \Sigma_k \hat{L_{i,j}}|_k.log_2(\frac{\hat{L_{i,j}}|_k}{\hat{L_{z,j}}|_k})$$

Method summary

Given, the joint probability density function of the data $X \sim p[f(\mathbf{X})\Theta]\ d\mathbf{X}$, effectively regard this as a *set* of marginal distributions $p[f(\mathbf{X}_{i,j})\Theta_{i,j}]\ d\mathbf{X}_{i,j}$ for all i and j (see Figure 12) and a *data* estimate of the covariance between them. Smooth by using distributions from the exponential family and summarise $p[f(\mathbf{X}_{i,j})\Theta_{i,j}]\ d\mathbf{X}_{i,j}$ by the height of the probability density function $L_{i,j}[f(X_{i,j})\hat{\Theta}_{i,j}]$ at the parameter estimates

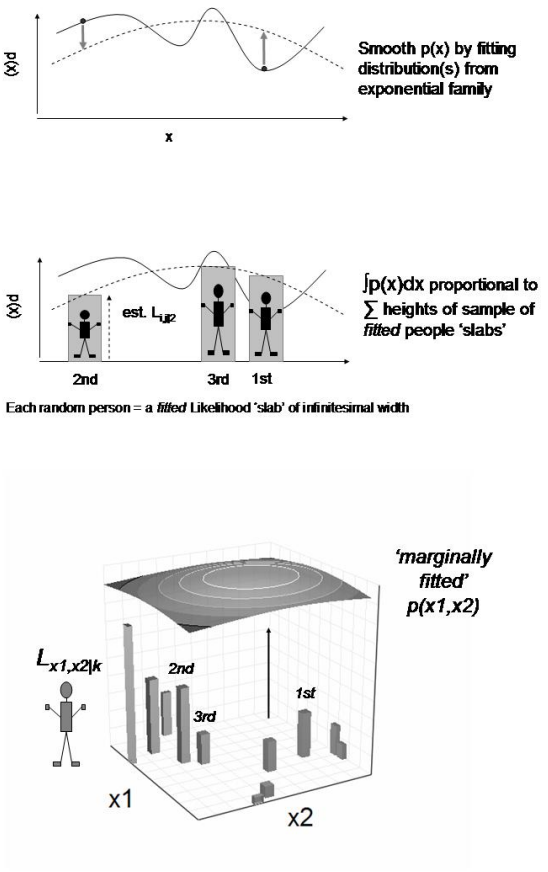


Figure 10. *Upper*: Fitting an exponential family distribution (to each margin) smooths the density. *Middle*: Now the Riemannian slabs are in the space of the fitted function. *Lower*:Each individual can now be looked up in the fitted joint bivariate probability density defined by the margins. Each bar is a subject. Likelihood height exaggerated and density with contours and shading for clarity. Note: examples here assume perfect knowledge of parameter estimates.

(i.e. the likelihood of the data given the estimates). Finally (for the example of the simple observed divergence or likelihood ratio in Figure 13) 'fold' the data with respect to the reference population z in X using the *relative* heights (aka the ratios of the likelihoods at their parameter estimates) individualised for the actual data observed (and then log transform - see Figure 14).

So, one can see that, if one recasts X as follows:-

$$X = \begin{bmatrix} x_{1,1,1} & x_{1,1,2} & \dots & x_{1,1,p} \\ x_{1,2,1} & x_{1,2,2} & \dots & x_{1,2,p} \\ \dots & \dots & \dots & \dots \\ x_{s,w,1} & x_{s,w,2} & \dots & x_{s,w,p} \\ \dots & \dots & \dots & \dots \\ x_{r,k,1} & x_{r,k,2} & \dots & x_{r,k,p} \end{bmatrix}$$

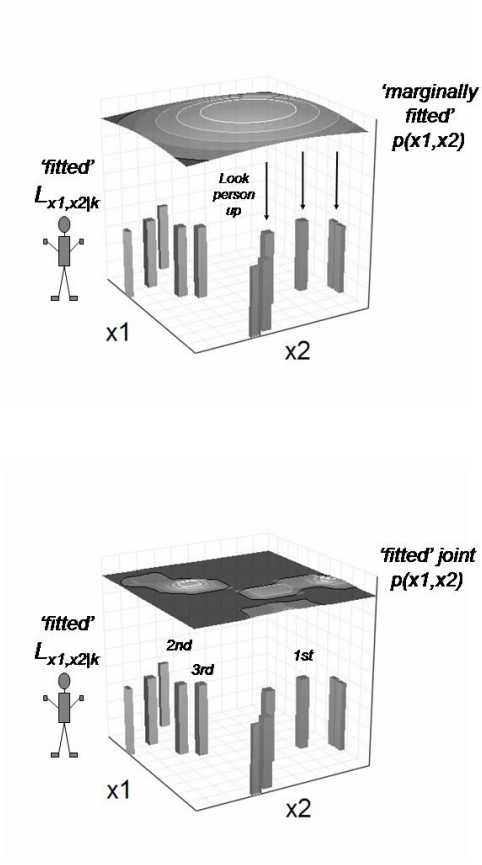


Figure 11. *Upper*: A likelihood value based upon the fitted density can be 'read-off'. *Lower*:The data covariance pattern (and subject 'clusters') remains of these individualised 'fitted' likelihood of a random sample of people - smoothing only non-linearly alters the relative heights of the likelihoods. Each bar is a subject. Likelihood height exaggerated and density with contours and shading for clarity. Note: examples here assumes perfect knowledge of parameter estimates.

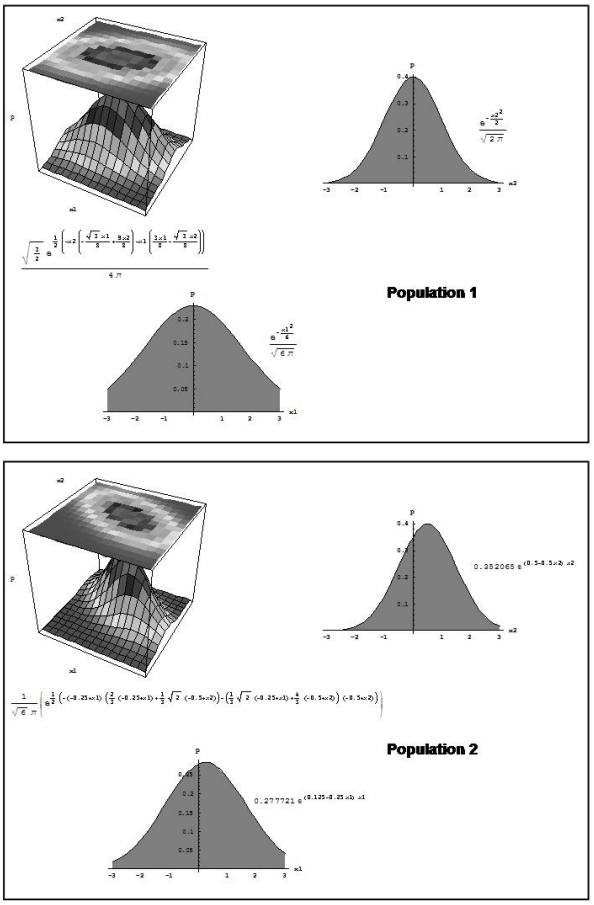


Figure 12. Basis of illustrative example. *Upper*: Population 1 - Bivariate Normal, expected values = (0,0) covariance matrix = $\begin{bmatrix} 3 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & 1 \end{bmatrix}$. Marginal distributions also shown; *Lower*: Population 2 ('reference') - Bivariate Normal, expected values = (0.25,0.5) covariance matrix = $\begin{bmatrix} 2 & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 1 \end{bmatrix}$. Marginal distributions also shown.

(pop. s=1...r, row w=1...k subjects, column c=1...p variables)

then $x_{s,w,c}$ is an observed data point. Now, under expectation, one could 'replace' each $x_{s,w,c}$ with the expected value for the sth population and the cth column i.e. $x_{s,E,c}$ - the typical value over the individuals for the cth variable in the sth population. The non-linear transformation using divergences is no different than this simple manipulation - $x_{s,w,c}$ is simply replaced by $d_{s,w,c}$ the divergence value for the sth population versus the zth reference population for the cth variate individualised for the wth subject. So, for population 1 (of n individuals) and reference population z (of m indviduals):-

$$X_1 = \begin{bmatrix} x_{1,1,1} & x_{1,1,2} & \dots & x_{1,1,p} \\ x_{1,2,1} & x_{1,2,2} & \dots & x_{1,2,p} \\ \dots & \dots & \dots & \dots \\ x_{1,n,1} & x_{1,n,2} & \dots & x_{1,n,p} \\ \Downarrow & \Downarrow & \dots & \Downarrow \\ \hat{\theta}_{1,.,1} & \hat{\theta}_{1,.,2} & \dots & \hat{\theta}_{1,.,p} \end{bmatrix},$$

$$X_z = \begin{bmatrix} x_{z,1,1} & x_{z,1,2} & \dots & x_{z,1,p} \\ x_{z,2,1} & x_{z,2,2} & \dots & x_{z,2,p} \\ \dots & \dots & \dots & \dots \\ x_{z,m,1} & x_{z,m,2} & \dots & x_{z,m,p} \\ \Downarrow & \Downarrow & \dots & \Downarrow \\ \hat{\theta}_{z,.,1} & \hat{\theta}_{z,.,2} & \dots & \hat{\theta}_{z,.,p} \end{bmatrix}$$

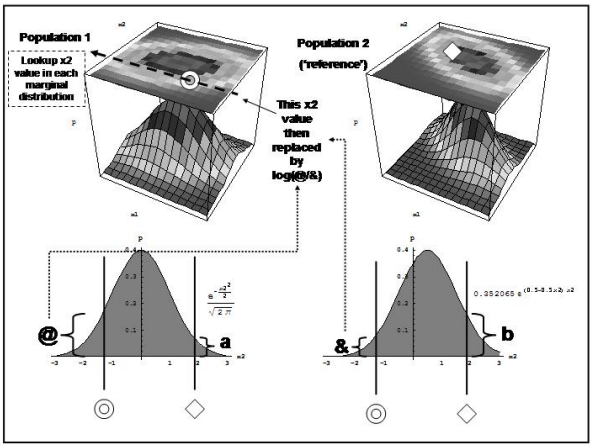


Figure 13. Bivariate illustration of kernel method - using marginal distributions for second variate x2, and observed log likelihood ratio as divergence from Figure 12. *Top*: Joint distributions for each population. 'Doughnut' subject belongs to population one, 'Diamond' subject belongs to population 2. When 'looked up' in each population's x2 marginal distribution, the 'Doughnut' subject has a higher likelihood (@) in population one and lower likelihood (&) in population two. The converse applies for the 'Diamond' subject. The x2 axis is then re-scaled for the 'Doughnut' subject using log(@/&), effectively non-linearly expanding or shrinking it in replacing the data point 'coordinates' exactly. The 'Diamond' subject would be re-scaled by log(a/b) as population 2 is the reference population.

Form (where | means 'evaluated at' and & means 'with') :-

$$D_1 = \left[\begin{array}{ll} d_{1,1,1}[\hat{\theta}_{1,.,1} \& \hat{\theta}_{z,.,1}] | x_{1,1,1} & , \\ d_{1,2,1}[\hat{\theta}_{1,.,1} \& \hat{\theta}_{z,.,1}] | x_{1,2,1} & , \\ \ldots & \\ d_{1,n,1}[\hat{\theta}_{1,.,1} \& \hat{\theta}_{z,.,1}] | x_{1,n,1} & , \end{array} \right.$$

$$\begin{array}{ll} d_{1,1,2}[\hat{\theta}_{1,.,2} \& \hat{\theta}_{z,.,2}] | x_{1,1,2} & , \ldots \\ d_{1,2,2}[\hat{\theta}_{1,.,2} \& \hat{\theta}_{z,.,2}] | x_{1,2,2} & , \ldots \\ \ldots & \\ d_{1,n,2}[\hat{\theta}_{1,.,2} \& \hat{\theta}_{z,.,2}] | x_{1,n,2} & , \ldots \end{array}$$

$$\left. \begin{array}{l} \ldots, d_{1,1,p}[\hat{\theta}_{1,.,p} \& \hat{\theta}_{z,.,p}] | x_{1,1,p} \\ \ldots, d_{1,2,p}[\hat{\theta}_{1,.,p} \& \hat{\theta}_{z,.,p}] | x_{1,2,p} \\ \ldots \\ \ldots, d_{1,n,p}[\hat{\theta}_{1,.,p} \& \hat{\theta}_{z,.,p}] | x_{1,n,p} \end{array} \right] = D_1$$

$$D_z = \left[\begin{array}{ll} d_{z,1,1}[\hat{\theta}_{1,.,1} \& \hat{\theta}_{z,.,1}] | x_{z,1,1} & , \\ d_{z,2,1}[\hat{\theta}_{1,.,1} \& \hat{\theta}_{z,.,1}] | x_{z,2,1} & , \\ \ldots & \\ d_{z,m,1}[\hat{\theta}_{1,.,1} \& \hat{\theta}_{z,.,1}] | x_{z,m,1} & , \end{array} \right.$$

$$\begin{array}{ll} d_{z,1,2}[\hat{\theta}_{1,.,2} \& \hat{\theta}_{z,.,2}] | x_{z,1,2} & , \ldots \\ d_{z,2,2}[\hat{\theta}_{1,.,2} \& \hat{\theta}_{z,.,2}] | x_{z,2,2} & , \ldots \\ \ldots & \\ d_{z,m,2}[\hat{\theta}_{1,.,2} \& \hat{\theta}_{z,.,2}] | x_{z,m,2} & , \ldots \end{array}$$

$$\left. \begin{array}{l} \ldots, d_{z,1,p}[\hat{\theta}_{1,.,p} \& \hat{\theta}_{z,.,p}] | x_{z,1,p} \\ \ldots, d_{z,2,p}[\hat{\theta}_{1,.,p} \& \hat{\theta}_{z,.,p}] | x_{z,2,p} \\ \ldots \\ \ldots, d_{z,m,p}[\hat{\theta}_{1,.,p} \& \hat{\theta}_{z,.,p}] | x_{z,m,p} \end{array} \right] = D_z$$

Note that if $\hat{\theta}$ are (jointly) sufficient statistics then any function (e.g. a divergence) is sufficient.

Note, also, (most importantly) that, in this methodology, *all* subjects get the same divergence d based upon $[\hat{\theta}_{1,.,1} \& \hat{\theta}_{z,.,1}]$ i.e. for the observed log likelihood ratio, *both* $d_{1,1,1}$ and $d_{z,1,1}$ get the *same* $log(\frac{\hat{L}_{i,j}}{\hat{L}_{z,j}})$ form evaluated at each individual - *not* $log(\frac{\hat{L}_{i,j}}{\hat{L}_{z,j}})$ for $d_{1,1,1}$ etc to $d_{1,n,p}$ and, the inverted $log(\frac{\hat{L}_{z,j}}{\hat{L}_{i,j}})$ for $d_{z,1,1}$ etc to $d_{z,m,p}$. In a case-control comparison, *all* individuals are thus treated as *if cases* for data replacement with divergences individualised for each subject (see Figure 15). That is, one replaces the data with the size of the contribution that the observation of the cth (marginal) variate on that wth person makes to the distinction between its population s and that of the reference population z evaluated at their parameter estimates (see Figure 14).One could consider this, in some sense, as 'fractionating' a χ^2.

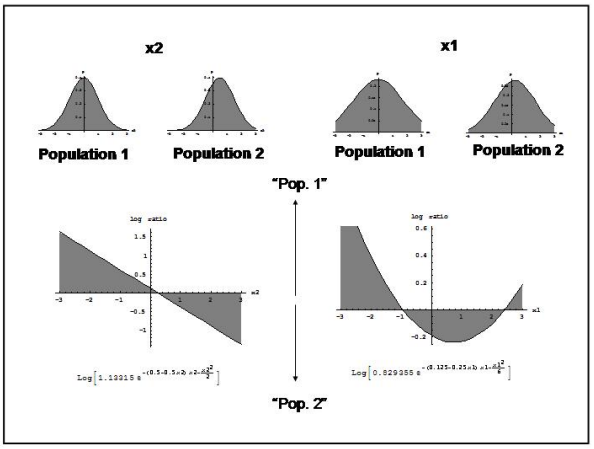


Figure 14. Illustration of kernel transformation functional from Figure 13. *Upper*: Using marginal distributions for second variate x2 (left) and first variate x1 (right) from bivariate normal example, observed log likelihood ratio $0.5*(\ln(\sigma_2^2)+\frac{(y-\mu_2)^2}{\sigma_2^2}-\ln(\sigma_1^2)-\frac{(y-\mu_1)^2}{\sigma_1^2})$ (where the subscripts indicate group and y the variate measured) as divergence (Bowman et al. 2006). As population 2 is the reference population, a positive functional value represents 'evidence' for population 1. Quadratic shape occurs since at large values the most dispersed marginal distribution is favoured. For uni-parameter exponential distributions (like the Binomial and Poisson) the linear form on left is typical (here for the normal distributed variates only because $\sigma_{1,x1} = \sigma_{2,x1}$ in this example.)

For example, replace $x_{s,w,c}$ with the expected log likelihood ratio estimate:

$$KL\hat{D}L^z_{s,c}|_w = \hat{L^z_{s,c}}|_w.log_2(\frac{\hat{L^z_{s,c}}|_w}{\hat{L^z_{z,c}}|_w})$$

which for normally distributed data is (half, Bowman et al. (2006)) of

$$\ln(\hat{\sigma_z^2}) + \frac{\hat{\sigma_s^2}}{\hat{\sigma_z^2}} + \frac{(\hat{\mu_z} - \hat{\mu_s})^2}{\hat{\sigma_z^2}} - \ln(\hat{\sigma_s^2}) - 1$$

By using the same functional form of the divergence measure across populations and retaining the indices c and w within s, the original *pattern* ϖ (but not scale and value) of any covariance structure shown by the data for each population is retained (it is simply non-linearly transformed). Accordingly, then form the overall covariance matrix of divergences Σ by estimation over the whole transformed data (see Figure 15).

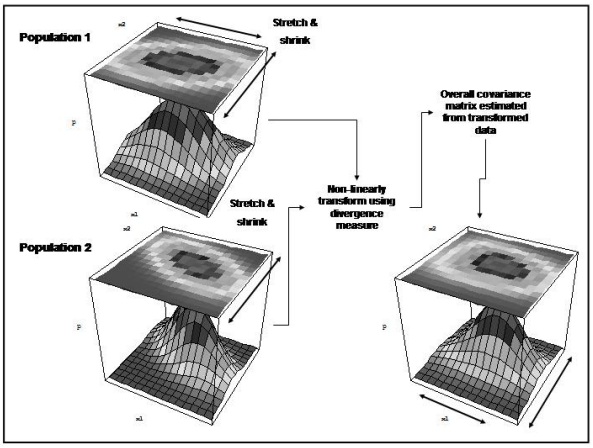
Figure 16 illustrates the flowchart for a two population example of the methodology.

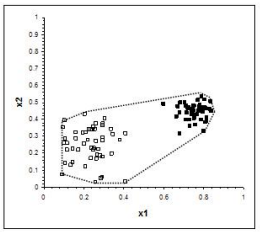
It is the covariation of data which captues phenomena of biological interest. This divergence method smooths, non-linearly transforming this and posing it in the space of the contrast (trait) of interest. Using PCA to decompose the non-linearly deformed data (Delrieu and Bowman 2006) of observed log likelihood divergences focuses on both within and between group variation, whilst for the expected measures it only focuses on the between group variation. Rather than SVD, non-negative matrix factorisation (Fogel et al. 2007) could be employed. Canonical variate analysis to maximise directions of between group variation given within group variation could be used on observed log likelihood divergences.

Confidence of estimates and Empirical null distributions

There may be variable confidence in estimates used in this way in divergences. Asymptotics of divergences has been recently summarised by **?**. Weights can be used appropriately as the Bayes factor (*lbf*) formulation explicitly allows for an *a priori* formulation. However, empirical data driven distributions off an alternative simpler way of forming inferences allowing for parameter estimation.

Given $\Theta = \begin{bmatrix} \theta \\ \psi \end{bmatrix}$ where ψ are nuisance parameters, θ could be considered as $\theta = \begin{bmatrix} \xi \\ \varpi \end{bmatrix}$ where ξ are the sufficient summary statistics making up the divergence measures (e.g. μ and σ^2 for the normal distribution) and ϖ is a *pattern* of covariances across the data variates. As Figure 15 indicates, divergences non-linearly stretch and shrink the variate axes and deform the existent *particular* covariation structure ϖ of the original data. That is - ξ, the relative sizes of marginal information losses of using the reference population z for each population i for each of the j variables (evaluated for each individual k), re-scales any fundamental ϖ.





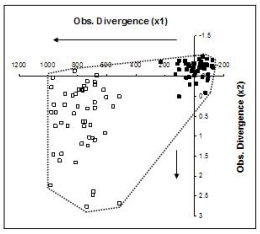


Figure 15. *Upper*: Bivariate illustration (from Figure 12) by which variate axes are non-linearly expanded or shrunk by the divergence measure deforming them (see Figures 13 and 14) into evidence measures. However, in replacing the 'co-ordinates' of the data point exactly it still retains the covariance structure *pattern* ϖ. Then, form an estimate of Σ over *all* the transformed data; *Lower* Simulated example with convex hulls (in dots). *Left*: Open squares 'case' population = simulated bivariate normal mean values = (0.25, 0.25) covariance matrix = $\begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$ Black squares (reference) 'control' population = simulated bivariate normal mean values = (0.75, 0.45) covariance matrix = $\begin{bmatrix} 0.05 & 0 \\ 0 & 0.05 \end{bmatrix}$ Overall estimated correlation matrix from data (Delrieu and Bowman 2006) for eigen decomposition = $\begin{bmatrix} 0.99 & 0.75 \\ 0.75 & 0.99 \end{bmatrix}$ *Right*: After application of observed log likelihood ratio functional (shrinking and swelling axes). Note: scales inverted. Very large 'information' on x1, small on x2 (1000 times on x1 compared to x2) - positive values (direction of arrows) indicate 'case-ness'. Divergences act as amplifiers of relevant distinction. Overall estimated correlation matrix from data (Delrieu and Bowman 2006) for eigen decomposition *now* = $\begin{bmatrix} 0.99 & 0.68 \\ 0.68 & 0.99 \end{bmatrix}$

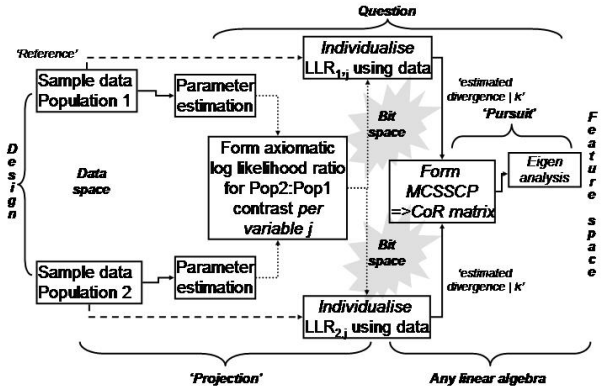


Figure 16. Flowchart of individualised likelihood divergences (Delrieu and Bowman 2006) - two group standardised variances example. MCSSCP = mean centred sums of squares and cross-products as estimate of Σ. If standardised variates this yields CoR = correlation matrix for eigen decomposition.

A simple permutation test of group status (e.g. case versus control attribution - Figure 15) for $\hat{\theta}$ effectively constructs a 'location' empirical null $p(\xi_0)$ for divergences (or derivations from them such as loadings) based upon $\xi = \hat{\xi}|\{\varpi = \hat{\varpi}\}$ i.e. given a covariation structure (conditionally) *fixed* at the original data estimate. This allows a univariate significance test of no effect (distinction) between groups for loadings etc (see http://taxonomy.delrieu.org). Asymptotic large sample distributional results under the null of no difference between groups (in location) evaluated at their MLEs are available for weighted sums of expected log bayes factors for various exponential distributions (see de Leon and Carrière (2003)).

Repeated drawing of a bootstrap sample of individuals (i.e. sampling with replacement) from *each* of the populations $i = 1...r$ followed by permutation of group status (but giving them the *original* un-permuted group data divergences values) effectively constructs an empirical null for divergences based upon $\xi = \hat{\xi}|\hat{\varpi}_w$ i.e. it adds a conditional 'wobble' or 'blur' into any 'location' null according to the estimation of *this* within group covariation structure. It little affects fundamental between group covariation.

Repeated drawing of a bootstrapped sample of individuals (i.e. sampling with replacement) from the *total* set of individuals at random, followed by giving them a population attribution (but giving them the *original* un-permuted group data divergences values) effectively constructs an empirical null for divergences based upon $\xi = \hat{\xi}|\hat{\varpi}_{w\&b}$ i.e. it adds a conditional 'wobble' or 'blur' into any 'location' null according to the estimation of *this* within and between group covariation structure.

Repeated drawing of a bootstrap sample of individuals (i.e. sampling with replacement) from *each* of the populations $i = 1...r$ at random and constructing *new* group divergences values effectively constructs a 'location' empirical null for divergences based upon simultaneously estimating $\hat{\xi}|\hat{\varpi}_w$ and $\hat{\varpi}_w|\hat{\xi}$ (i.e. it adds a *joint* 'wobble' or 'blur' to both within group estimates in any null). It little affects fundamental between group covariation.

Repeated drawing of a bootstrap sample of individuals (i.e. sampling with replacement) from the *total* set of individuals at random, followed by giving them a population attribution and constructing *new* group divergences values effectively constructs a 'location' empirical null for divergences based upon simultaneously estimating $\hat{\xi}|\hat{\varpi}_{w\&b}$ and $\hat{\varpi}_{w\&b}|\hat{\xi}$ (i.e. it adds a *joint* 'wobble' or 'blur' to both within and between group estimates of location and covariation structure in any null).

Column-wise permutation (or bootstrap sampling) of variates retains the *typical* covariate structure of the data ϖ but adds stochasticity on the estimate of that typical structure i.e. producing estimates of $\hat{\varpi}_0$ given *these* variates (or from a *population* of these variates, respectively). Such shuffling within each population addresses second order stochasticity within populations; co-ordinated shuffling over all populations addresses 'blur' in second order relations between populations. Disco-ordinated shuffling offers 'blur' in second order relations within and between populations. Original or recalculated ξ values can be used as desired.

Taking a simulated random field as data (i.e. an *unstructured* ϖ) offers the basis for tests for ξ (fixed, conditional on or joint with $\hat{\varpi}$ as desired - see above) versus assuming no particular covariation structure at all - this is effectively assuming some form of 'random' network. A simulated uniform field would mirror a fully-connected network of similar edge strengths (which is easier for humans to see differences to than a random one) but is biologically unrealistic. Null matrices generated from a historically grounded random walk would perhaps be more appropriate. For simultaneous analyses of variates of different biological scale or system 'niveau' (see Delrieu and Bowman (2007)), perhaps self-similar nulls should be used.

Whilst the intention in using null distributions for testing $\hat{\xi}$ is not usually to also test the 'significance' of $\hat{\varpi}$, SVD effectively triages data along latent axes of maximum variation, so its basis does detect such changes as ξ rescales ϖ. Similarly, as there is a mapping of covariance matrices to a network representation via a (thresholded) walk laplacian (Bowman and Delrieu 2008), the occurrence of variate 'nodes' as well as the existence or strength of 'edges'

between nodes will vary with $\hat{\varpi}$ as well as with the non-linearly scaling sizes of $\hat{\xi}$. Assumptions of exchangeability (Jaynes 2003) matter and in testing subjective options abound!