

Whole genome scan algebra and smoothing

Christakis Charalambous¹, Olivier Delrieu² and Clive Bowman^{3*}

¹University of Cyprus, Nicosia, Cyprus; ² CEPH, Paris, France and ³ Genetics, GlaxoSmithKline R&D, Greenford Road, Greenford, London UB6 0HE, UK

Abstract

The algebra of comparing genome scans within and between studies with distributional-based divergences is presented together with a mathematical theory for smoothing them. A small study exemplar shows that standardisation by an inverse of an approximation to, or an inverse of only large scale, control linkage disequilibrium can be sufficient.

Key words:- bio-marker, supervised dimensional reduction, divergences, filtering, gene, log likelihood ratio, multivariate method, SNP, visualisation, exponential basis function, Bayes Factor, single value decomposition, LD.

Introduction: High-dimension low-sample-size whole genome scans (of $p \sim O[10^6]$ 'components' for $n \sim O[10^{2-3}]$ samples) are increasingly being used in the development of drugs (Roses, 2004). Such SNP data can be expressed in terms of (individualised) class divergences (Jardine and Sibson, 1971) – natural information-content basis functions (*bf*) for the comparison of individuals or groups in the co-analysis or simultaneous visualisation (i.e. the filtering) of multiple bio-marker data (Delrieu and Bowman, 2005). These *bf* have simple convenient closed forms for data from the multi-parameter exponential family of distributions (Delrieu and Bowman, 2007) and can encompass the response (Bowman *et al.*, 2006) and covariates (Delrieu and Bowman, 2007). The simplest divergence is the log-likelihood (log Bayes Factor or *lbf*). Screening by over-complete *bf* projection (i.e. on every data axis and every interaction axis (Delrieu and Bowman, 2006b) ensures a logical polythetic 'Manhattan grand tour' of the shape or manifold of any multidimensional data cloud relevant to the question being asked (Delrieu and Bowman, 2006a).

Nature is such that so-called 'high-throughput' biological data, often multi-colinear, is internally redundant showing permutation-invariant and label-dependent sparsity due to complexity constraints i.e. there are truly low degrees of freedom in the latent data-generation process ("*Raffiniert ist der Herrgott aber boshaft ist er nicht*" – Einstein 1921). Given this, despite indeterminacy (since $p \gg n$) and computational difficulties at large p , eigen analysis of the correlations over the combined dictionaries of such 'component'-wise divergences within a study (<http://taxonomy.delrieu.org>) can yield useful phenomenologically *insightful* sets of multi-scale condensed features under a Pareto hypothesis (i.e. a 'hypothesis of effect sparsity'; Pirmohamed *et al.*, 2007). Covariances of divergences can be decomposed similarly, but are scale dependent and focus on discovering 'significant' sets of low dimensional condensed *predictive* features. Subjectively providing a definition for a class (*induction*) is not the same task as objectively identifying a member of a class (*deduction*) (Van Regenmortel 2006). Both abstraction and diagnosis often show approximate eigenstructure sparsity (spiked eigenvalue scree plots). Sparsity ensures that classification (say by naive Bayes) performs well – mitigating Bellman's (Bellman, 1961) apparent data-space 'curse of dimensionality' (Clarke *et al.*, 2008) and the

dangers of noise accumulation at large p (Fan and Fan, 2007). Eigen vectors can be structurally inconsistent for low n , but increasing p can sharpen pre-existing data definition in this empirical supervised dimensional reduction. Regularisation through adding other compressed and structured *bf* 'components' such as ontologies (Delrieu and Bowman, 2006a), distances, paths, manifolds, hierarchies, unlabeled data covariances/correlations, prior beliefs, complexity measures etc can be done to avoid over-fitting. Dummy variables can be added to exemplify hypotheses. Condensed feature sets encapsulate genotype and phenotype integration (Pigliucci and Preston, 2004).

Algebra: The exact biological meaning of whole genome scan algebraic manipulation varies according to the divergence measure (Delrieu and Bowman, 2007) and profile 'components' used. However in these transformed co-ordinate systems, simple within study moment-based summaries of divergences highlight various data features (Table 1). In turn, the subtraction of these moments also has simple useful interpretations whether between individuals or at the group level (Table 1). The distribution of between individual pair-wise differences also yields insights (Figure 1 Top Left, Murtagh 2007). Higher profile moments (such as variance, skewness and kurtosis) are indicators of innovativeness and non-unanimity (or conversely unanimity) – that is they indicate the degree and patterns of controversy of genetic signals within a profile (or conversely the degree of their consensus). They statistically describe the shape of polygenicity. Since divergences are a universal appropriate basis (or 'metric' *sensu lato*) for comparisons, pooling of phenotypes within a study or pooling of raw data from multiple studies followed by multi-phenotype dissection is feasible. Pirmohamed *et al.* (2007) gives an exposition of a within-study multiple phenotype analysis. Between-study phenotype subtraction using summary measures of divergences can be similarly made with useful interpretations (Table 2). Weighted averaging of these 'Table 2' measures over studies is analogous to meta-analysis.

Smoothing: Linkage disequilibrium (inherited co-occurrence of markers more than by chance alone due to selection or founder effects) can be thought of as a nuisance parameter in the comparison of genetic signals. Signal covariation subsumes both LD and stochastic sampling effects - so why not take a log Bayes Factor (*lbf*) divergence vector A measuring the association between markers and a group difference and multiply it by the 'inverse' of the square root of its covariance matrix $B^{-\frac{1}{2}}$ to yield a unit-less LD-adjusted standardisation of the *lbf* profile? Such inverse covariance standardisation will re-scale and smooth genetic profiles by removing LD.

Singular value decomposition of the original data encompasses all the information available in its covariance matrix, hence one can project (display) the case data say in the column space of just the controls. Since, let the measured features of the *ith* person in a case-control study be:

$$\tilde{x}^{(i)} = \begin{bmatrix} \tilde{x}_1^{(i)} \\ \tilde{x}_2^{(i)} \\ \dots \\ \tilde{x}_N^{(i)} \end{bmatrix} \text{ where } i \text{ denotes the person, } p \text{ the number of persons, } N \text{ the number of SNPs,}$$

$\tilde{x}_j^{(i)}$ the *lbf* for the *j*th SNP corresponding to person *i*. Then the measured feature matrix is $\tilde{X} = [\tilde{x}^{(1)} \tilde{x}^{(2)} \dots \tilde{x}^{(p)}]$. $p \ll N$ and the columns of \tilde{X} (aka individuals) are linearly independent (i.e. $\text{rank}(\tilde{X})=p$). The row mean of \tilde{X} is $\tilde{\mu} = \frac{1}{p} \sum_{j=1}^p \tilde{x}^{(j)}$. The translated measured feature vectors are $x^{(i)} = \tilde{x}^{(i)} - \tilde{\mu}$, $i = 1, 2, \dots, p$. Leading to a measured feature matrix $X = [x^{(1)} x^{(2)} \dots x^{(p)}]$. $x \in R^N$. The covariance matrix of X is $S = E[XX^T]$ estimated herein by $\frac{1}{p-1} XX^T$. The square root of S is $Q_h = S^{1/2}$ i.e. $S = (Q_h Q_h)$. The rank of X is $(p - 1)$. Matrix S is positive semidefinite and its column space is the same as the column space of X . The column

Phenotype	Subject	SNPs					Summary (over p)
Case	1	$\alpha_{1,1,1}$	$\alpha_{1,1,p}$	Measure $M_{1,1}, V_{1,1}$
Case	2	Measure $M_{1,2}, V_{1,2}$
Case
Case	n	$\alpha_{1,n,1}$	$\alpha_{1,n,p}$	Measure $M_{1,n}, V_{1,n}$
	E or Var (over n)	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$...	$a_{1,p}$	d_1 or d_1^*
Cntl	1	$\alpha_{2,1,1}$	$\alpha_{2,1,p}$	Measure $M_{2,1}, V_{2,1}$
Cntl	2	Measure $M_{2,2}, V_{2,2}$
Cntl
Cntl	m	$\alpha_{2,m,1}$	$\alpha_{2,m,p}$	Measure $M_{2,m}, V_{2,m}$
	E or Var (over m)	$b_{2,1}$	$b_{2,2}$	$b_{2,3}$...	$b_{2,p}$	e_1 or e_1^*
Overall	E or Var (over n+m)	c_1	c_2	c_3	...	c_p	f_1 or f_1^*

Table 1: Example of within-study summaries of divergences for a case-control study. $\alpha_{1,1,1}$ = divergence value instantiated for 1st group, 1st subject and 1st SNP ... etc. $\alpha_{1,1,1}$ is the information the first SNP carried by the first subject in group 1 has relevant to the comparison posed, and so on. $a_{1,1}$ (or $b_{2,1}$) is the expected 'typical' value (E) or the variability (Var) of information the first SNP carried by group 1 (or 2) has pertinent to the posed contrast etc. High typicality suggests possible useful features. High variability suggests subject heterogeneity. $(a_{1,1} - b_{2,1})$ is the difference in these between the groups - high typical values indicating important features (cf. 'signals') or large differences in the within group variability pointing to possible sporadic outlier subjects. c_1 is a suitably scaled total of these moments for the study. The marginal measures M and V as per person profile moment summaries are explained in Delrieu and Bowman (2005, 2006a), and Pirmohamed *et al.* (2007). $(\alpha_{i,j,1} - \alpha_{k,l,1})$ is the pair-wise difference in divergences for the first SNP between the jth individual of group i and the lth individual of group k (a j (of group i) to l (of group k) inter-individual *distance*, Figure 1). Starred variances (*) can have two values (and two interpretations - see // below) depending upon the order of calculation - 'over p then over n (or m or n+m)' versus 'over n (or m or n+m) then over p'. Summaries d and e represent group typical profiles and their variability. High typicalities suggest an informative study. Low variabilities indicate a large number of relevant SNPs (or uniform typical profiles // homogeneous profiles); high variabilities indicate a small number of relevant SNPs (or 'spiky' typical profiles // heterogenous profiles). (d-e) is the difference in group profiles or variabilities. A large typical difference suggest a highly informative study for the question posed. A large difference in variability points to sporadic outlier subjects or possible unequal 'component' covariance structure between the groups. Suitably scaled f characterises the overall study profile (discriminative *size* of the study) and its variability .

Phenotype	Subject	SNPs					Summary (over p or q)
Case X	E or Var (over n)	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$...	$a_{1,p}$	d_1 or d_1^*
Cntl X	E or Var (over m)	$b_{2,1}$	$b_{2,2}$	$b_{2,3}$...	$b_{2,p}$	e_1 or e_1^*
Overall X	E or Var (over $n+m$)	c_1	c_2	c_3	...	c_p	f_1 or f_1^*
Case Y	E or Var (over k)	$A_{1,1}$	$A_{1,2}$	$A_{1,3}$	$A_{1,q}$		D_1 or D_1^*
Cntl Y	E or Var (over l)	$B_{2,1}$	$B_{2,2}$	$B_{2,3}$	$B_{2,q}$		E_1 or E_1^*
Overall Y	E or Var (over $k+l$)	C_1	C_2	C_3	C_q		F_1 or F_1^*

Table 2: Between study summaries of divergences (Phenotype X and Phenotype Y, $p > q$ in this example). Starred variances (*) can have two values depending upon order of calculation (see Table 1). ($A_{1,1} - a_{1,1}$) and so on, indicates any differences in typicality or variability between the case populations of phenotypes Y and X (adjusted for the control genetic basis of each study) subject to any systematic genetic offset between the control populations for the first SNP. The latter would allow say, the discovery of pre-disposing genes for ocular SJS in Japanese (Ueta *et al.*, 2007a, 2007b) (=phenotype Y) versus general SJS in Chinese (Chung *et al.*, 2004) (=phenotype X) given an understanding of ethnic differences in allele frequencies in healthy individuals. Given comparable controls between phenotype X and Y, ($C_1 - c_1$) measures the absolute differences in typicality or variability between the case populations of phenotypes Y and X (cf. phenotype subtraction over genetic features - see Pirmohamed *et al.* (2007) page 1691 Figure S15). Differences between profile moments ($D_1 - d_1$), and ($F_1 - f_1$) follow similarly. For $X = Y$, ($\frac{a_{1,1}+A_{1,1}}{2}$) or ($\frac{d_1+D_1}{2}$) represent meta-analytical summaries across studies.

space of Q_h is the same as the column space of S . Now, consider the system of equations: $Q_h y = x$. This system is solvable if and only if x is in the column space of S and hence of X . In this case, there are an infinite number of solutions but the solution of minimum length is unique. If x is *not* in the column space of X then we look for a solution minimising the least squares error function $LSEF = \|Q_h y - x\|_2$. Using singular value decomposition (SVD) we can express X as $X = V \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} U^T$ where V is a $N \times N$ orthogonal matrix (a set of orthonormal 'output' basis vector directions for X), and U is a $p \times p$ orthogonal matrix (a set of orthonormal 'input' or 'analysing' basis vector directions for X). $\Sigma = diag(\sigma_i)$ where $\sigma_1 \geq \sigma_2 \dots \geq \sigma_r$, the non-zero eigenvalues of $X^T X$ (or equivalently $X X^T$) being arranged in descending order (a set of scalar 'gain controls' by which each corresponding input is multiplied to give a corresponding output). The columns of U are the corresponding eigenvectors of $X^T X$ (aka right singular vectors). The columns of V are the corresponding eigenvectors of $X X^T$ (aka left singular vectors). Then, $S = V \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} V^T$, $Q_h = V \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^T$ and the value of y that minimises LSEF is $y^* = V \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{bmatrix} V^T x$. Define V_1 as the first r columns of V . The rank of $V_1 = r \ll N$. $\delta^* = \Sigma^{-1} V_1^T x$ are the co-ordinates of y^* with respect to eigenvectors v^1, v^2, \dots, v^r for plotting. Defining the pseudo-inverse (Rao, 1973) of Q_h as $Q_h^+ = V \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{bmatrix} V^T$ then $y^* = Q_h^+ x$ and this is the min(LSEF) smallest 2-norm solution. The vector $x_c^+ = Q_h Q_h^+ x$ is the orthogonal projection of x onto the column space of Q_h and hence to the column space of X , and, the vector $r = x - Q_h y^* = (I - Q_h Q_h^+) x$ is the projection of x onto the subspace orthogonal to the column space of Q_h (and hence orthogonal to the column space of X). Then $X^T r = 0$ and $x = x_c^+ + r$ (Figure 1 Top Right).

If we define the columns of the measured feature matrix to correspond only to the *controls*, then solve as above and apply to all subjects, the transformation from $x \rightarrow y$ i.e. $y = Q_h^+ x$ has the effect of 'control' smoothing (Figure 1 Middle Left) the 'input' signals x due to the fact that Q_h^+

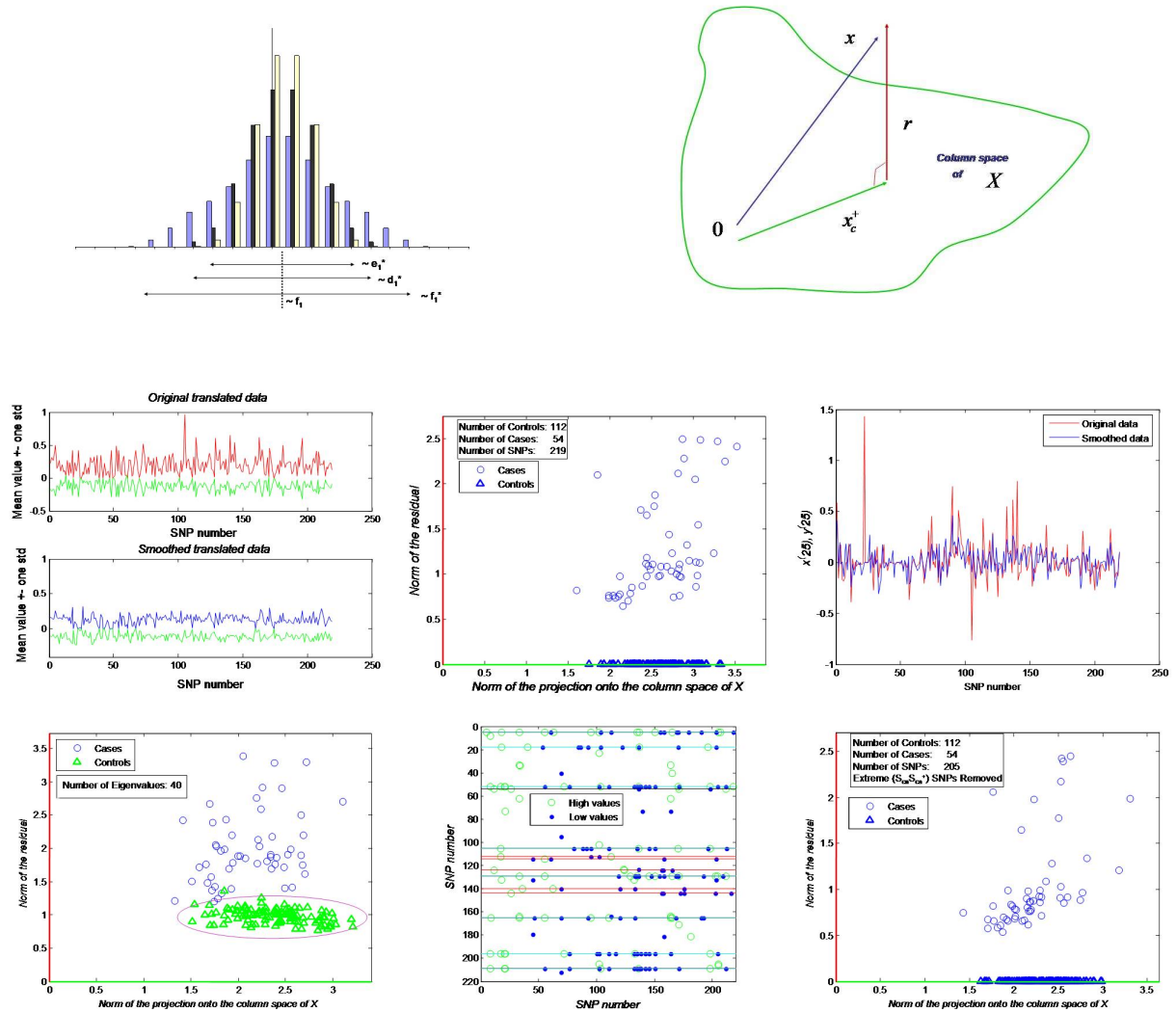


Figure 1: Top Left: Histogram of inter-individual pair-wise difference in Measure M from *lbf*s for cases, controls and overall in Pirmohamed *et al.* (2007). White bars = Cases, Black bars = Controls, Grey bars = Overall. Y axis at zero. \sim = 'indicative of'. Controls are more compact - indicative of smaller e_1^* in Table 1 Upper. Cases are more dispersed - indicative of larger d_1^* in Table 1 Upper. Small size of study (Pirmohamed *et al.*, 2007) mirrored by small size of f_1 (dotted). Increased spread overall indicative of f_1^* and presence of case-control genetic differences. Top Right: Geometric illustration that x_c^+ is in the column space of X and the residual r is orthogonal to the column space of X . Middle Left: Profile of study divergence over all subjects for each SNP (c_i in Table 1 Upper) in Pirmohamed *et al.* (2007) before and after control subspace smoothing using all eigenvalues. Upper trace in each panel is mean over individuals, lower trace in each panel is mean minus one standard deviation over individuals (i.e. $E[c_i] - \sqrt{Var[c_i]}$). SNP order along chromosomes. Middle Centre: Plot of the norms of the projection of the cases and controls onto the *control* column space of X and onto the subspace orthogonal to X for Pirmohamed *et al.* (2007) showing case heterogeneity having removed control LD. Controls on the x axis *are* in the control subspace. Vertical axis represents an LD-adjusted genetic distance for each individual. Middle Right: *lbf* profile for case 25 before and after control $LD^{-1/2}$ smoothing using all eigenvalues. Noticeably 'spiky' profile is original data. Note positive concerted smoothed signal for SNPs 80-100 in this individual. Bottom Left: Plot of the norms of the projection of the cases and controls onto the *control* column space of X and onto the subspace orthogonal to X for Pirmohamed *et al.* (2007) for first 40 eigenvalues/vectors only (\equiv 85% of variation). Note still separation of cases and controls. Controls (circled) are no longer quite planar in the original control subspace. Bottom Centre: Elements of high and low values of the matrix $S_{cn}S_{ca}^+$ joined by horizontal lines. Only 14 SNPs show appreciable differences in covariation between the cases and controls. Bottom Right: Plot of the norms of the projection of the cases and controls onto the *control* column space of X and onto the subspace orthogonal to X for Pirmohamed *et al.* (2007) after removal of 14 SNPs with subjectively extreme values of $S_{cn}S_{ca}^+$. LD disparities between the groups have little impact on case-control distinction.

tries to equalize the variations of the transformed data along orthogonal directions of maximum variance (\equiv removal of *control* LD and sampling stochasticity). The length of the residual $\|r\|_2$ for any *case* vector $x \in R^N$ can tell us how far the particular case is from the *control* subspace (i.e. a control LD-adjusted genetic distance, Figure 1 Middle Centre). Individual profiles can be smoothed (see case 25 in Figure 1 Middle Right) to remove the pattern of control LD. We can even approximate X by $X^{(k)}$ where $X^{(k)} = \sum_{i=1}^k \sigma_i \nu^{(i)} u^{(i)T}$ and $k \leq \text{rank}(X)$. Then $\text{Min}_B \|X - B\|_2 = \|X - X^{(k)}\|_2 = \sigma_{k+1}$, $\text{rank}(B)=k$ and $\|X\|_2 = \text{Sup}_{z \neq 0} \frac{\|Xz\|_2}{\|z\|_2}$. Then in the reduced (k) subspace, $S^{(k)} = V_1^{(k)} \Sigma_k^2 V_1^{(k)T}$, $Q_h^{(k)} = V_1^{(k)} \Sigma_k V_1^{(k)T}$, $Q_h^{(k)+} = V_1^{(k)} \Sigma_k^{-1} V_1^{(k)T}$ and $y^{(k)*} = Q_h^{(k)+} x$. Using such a subset of eigenvalues and eigenvectors (Figure 1 Bottom Left) for projection shows that inter-group subspace differences do not depend greatly upon fine distinctions captured by small 'noise' components. Smoothing by modest control LD decomposition (i.e. approximate standardisation) can be sufficient. Of course marker covariation (LD) may be different between cases and controls. So, define $\Lambda = S_{cn} S_{ca}^+$ where S_{cn} is the marker covariance matrix for controls and S_{ca}^+ is the (generalised) inverse of the marker covariance matrix for cases. The high and low values in Λ will, in a unit-less way, subjectively assess any difference in second-order patterns. This is where one would expect a strong case-control signal. For study Pirmohamed *et al.* (2007), Figure 1 Bottom Centre shows that only 14 out of the 219 SNPs have high or low extreme values of Λ . Removing these (as a 'strawman') and re-projecting on the control sub-space shows that the essential separation in subspaces is retained i.e. there is little appreciable difference to the case-control subspace distinction (Figure 1 Bottom Right). These 14 SNPs are not important for smoothing. The subspace distinction is everywhere in the profile. Standardisation by adjusting for the large scale control LD pattern can be sufficient.

Acknowledgement: We could not have developed these ideas without the support of Lefkos Middleton (GSK Collaboration G3333) and the enduring personal encouragement of Allen Roses.

References

- Bellman, R. (1961) *Adaptive control processes: A guided tour.* (A RAND Corporation Research Study), Princeton University Press, XVI
- 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
- Chung, W., Hung, S., Hong, H., Hsieh, M., Yang, L., Ho, H., Wu, J., and Chen, Y. (2004). Medical genetics: A marker for Stevens-Johnson syndrome *Nature*, **428** (6982), 486
- Clarke, R., Ransom, H., Wang A., Xuan, J., Liu, M., Gehan, E., and Wang, Y. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, *Nature Reviews Cancer*, **8**, 37-49
- Delrieu, O., and Bowman, C. (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, 21-24
- Delrieu, O., and Bowman, C. (2006a). Visualising gene determinants of disease in drug discovery *Pharmacogenomics*, **7**(3), 311-329

- Delrieu, O., and Bowman, C. (2006b). Visualisation of gene by gene interactions in pharmacogenetics *International Congress Of Human Genetics, Brisbane Australia, 6-11th August 2006* (poster)
- Delrieu, O., and Bowman, C. (2007). On using the correlations of divergences. In: S. Barber, P.D. Baxter, and K.V. Mardia (Eds), *Systems Biology and Statistical Bioinformatics*. University of Leeds, 27-35
- Fan, J., and Fan, Y. (2007). High dimensional classification using features annealed independence rules. *The Annals of Statistics*, to appear
- Jardine, N., and Sibson, R. (1971). *Mathematical Taxonomy*, John Wiley
- Murtagh, F. (2007). The remarkable simplicity of very high dimensional data: Application to model-based clustering. *SIAM Journal on Scientific Computing*, submitted
- Pigliucci, M., and Preston, K. (2004) *Phenotypic integration: Studying the ecology and evolution of complex phenotypes*. Oxford University Press
- 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
- Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, Wiley
- Roses, A.D. (2004). Pharmacogenetics and Drug Development: The path to safer and more effective drugs. *Nature Reviews Genetics*, **5**(9), 643-655
- Ueta, M., Sotozono, C., Tokunaga, K., Yabe, T., and Kinoshita, S. (2007a). Strong Association Between HLA-A*0206 and Stevens-Johnson Syndrome in the Japanese, *American Journal of Ophthalmology*, **143**(2), 367-8
- Ueta, M., Sotozono, C., Inatomi, T., Kojima, K., Hamuro, J., and Kinoshita, S. (2007b). Association of IL4R polymorphisms with Stevens-Johnson syndrome *J. Allergy Clin. Immunol* (in press doi:10.1016/j.jaci.2007.07.048)
- Van Regenmortel, M.V.H. (2006). Virologists, taxonomy and the demands of logic, *Arch. Virol.*, **151** (7), 1251-1255