

Taming variability: advances in the statistical analysis of complex data¹

Emmanuel Paradis² & Vincent Debat

November 28, 2002

Abstract

Recent advances in statistical and computational methods have resulted in an increased availability of approaches to analyse complex data with possibly several levels of variability. These methods handle, with a parametric approach, nonindependence, variance heterogeneity, and non-normality, making them applicable to a wide range of actual situations. Thus, variability, in its different forms, can be considered as the subject of interest, and not as a nuisance. This is an invitation to ecologists and evolutionists to give more emphasis on parameter estimation in data analysis than is usually done. An understanding of the methods reviewed here is likely to give insights in current issues in ecology and evolutionary biology.

1 Introduction

Variation, variability, and heterogeneity are common words in the ecological and evolutionary literature. Many current issues are centered the variable nature of important biological quantities. What is the relationship between environmental variability and life-history

¹© 2002 Emmanuel Paradis & Vincent Debat

Permission is granted to make and distribute copies, either in part or in full and in any language, of this document on any support provided the above copyright notice is included in all copies. Permission is granted to translate this document, either in part or in full, in any language provided the above copyright notice is included.

²Laboratoire de Paléontologie, Paléobiologie & Phylogénie, Institut des Sciences de l'Évolution, Université Montpellier II, F-34095 Montpellier cédex 05, France; e-mail: paradis@isem.univ-montp2.fr

strategies? How does spatial heterogeneity affect persistence of endangered species? Does environmental variability synchronize local populations? What are the effects of temporal fluctuations in resources on genetic variance components? How do mutation rates vary in DNA sequences? These are a few examples of issues where variability should be treated as the subject of interest rather than a nuisance in data analysis.

The statistical methods commonly used by ecologists focus on the mean tendency in the data through a linear model (e.g., linear regression, analysis of variance). The linear model assumes that variance is homogeneous, meaning variance heterogeneity is not formally appraised. This appears to be in contrast with the importance of variability in ecology and evolutionary biology. The discrepancy is understandable since much of the progress on statistical methods that aim to analyse variability *per se* has been very recent. This progress has led to new developments in computational methods and software, resulting in the availability of these methods to non-specialists in a wide range of situations. Also, new graphical methods have brought fresh perspectives in exploratory analyses.

There is a wide range of situations where the assumption of homogeneous variance does not hold: for instance, it is common to observe a heterogeneous dispersion of the data in scatterplots between macro-ecological relationships such as geographic range, abundance, and body size [7,8,11]. Another assumption which is often violated is that of independent observations. Indeed, data are often collected repeatedly on the same individuals, or on individuals that may be genetically related, or that may interact in the field, and so on. This phenomenon is called *clustering* of observations, and results in non-null covariances among observations which is typically modelled with a correlation matrix (Appendix 1). A slightly more complicated situation is when a parameter is better viewed as varying randomly rather than systematically (Fig. 1). Finally, OVERDISPERSION is a common form of variability when analysing count data.

There is a duality between considering variability as a nuisance or as the subject of interest. Whether a student is in one situation or the other actually depends on the context. From a methodological point of view, there is not a great difference between the two situations, and the methods reviewed here could be applied in both.

A common and well-known consequence of the failure to take the sources of variability described above into account is an inflated type

I error rate [16,22]. There exist some methods to correct for this in traditional statistical methods, but modelling it explicitly has been shown to be a much more efficient and flexible approach in several situations, such as the analysis of linkage in quantitative trait loci [54], modelling growth curves [27], sex ratio analysis [17], meta-analysis [38,47], or analysis of repeated measures [22].

2 Joint mean-variance modelling

Quite different phenomena, such as heteroscedasticity or overdispersion, can be handled with the joint modelling of mean and variance. This can be done by using the usual linear model for the mean $Y = \beta X$, which implies that we have for the i th observation $y_i = \beta x_i + \epsilon_i$, and another model for the variance $\text{Var}(\epsilon_i) = \exp(\lambda z_i)$ instead of the usual assumption of $\text{Var}(\epsilon_i)$ constant [1]. A similar approach can be used with non-normal data where in place of a standard linear model for Y a GENERALIZED LINEAR MODEL (GLM) will be used [18,19].

Joint mean-variance models have several virtues. They can easily be fitted with standard softwares providing the procedures to fit linear models and GLMs can be called repeatedly. Heterogeneous variances can readily be interpreted biologically since they may be related to their covariates, and their statistical significance can be tested (e.g. by testing $H_0: \lambda = 0$). As mentioned above, this approach can include several special cases, and thus appears as a simple and flexible method, intermediate between the corrections of the linear model, and the more complex (but more general) mixed-effects models. Surprisingly, joint mean-variance modelling does not seem to have been considered by ecologists though there are several potential applications, such as the analysis of heterogeneity in demographic parameters in relation to population density [2,33].

3 Mixed-effects models

The mixed-effects approach covers a wide range of models (Appendix 2). Typically, a mixed-effects model (or ‘mixed model’ in short) includes both fixed and random effects. A fixed effect is determined by a coefficient which is considered constant, whereas the coefficient of a random effect is a random variable. From this formulation, one could be interested in estimating simply the variance of these random parameters,

and this is known as the estimation of *variance components*. It is sometimes of interest to “estimate” the random effects themselves, that is the value of these random parameters that has been realized for each observation. The random effects cannot be estimated like a parameter, and the term “predicted” is used: this is done with a method called BEST LINEAR UNBIASED PREDICTOR.

Whether an effect should be considered as fixed or random depends on the context, and a matter of judgement for the experimenter given his knowledge of the data. Considering an effect as random may greatly help to reduce the number of parameters in a model, from a possibly large number of fixed effects to a single variance component. For instance, imagine a data set collected on a large number of individuals where the interest is in a relation between a morphological and an environmental variable. There are good reasons to suppose this relation is influenced by effects at the level of the individual such as the genotype, or the local environment. Considering individual as a fixed (categorical) effect would require including a very large number of parameters in the model (the number of individuals minus one). The alternative is to consider individual as a random effect requiring the estimation of only a single variance component. Conceptually, this is similar to consider this random effect to be the result of the many potential effects at the level of the individual. From a practical perspective, the tests on the other (fixed) effects will not be biased by this inter-individual variation. Furthermore, if several measures were made on the same individuals, it is possible to take the inherent correlation between them into account [6,35,39,42].

Mixed models have been used for several years in the analysis of animal breeding experiments with the use of ‘animal models’ where individuals are treated as random effects, and genetic variances are estimated with variance components estimators. The animal model approach has been applied successfully in a wild bird population [25,26].

Mixed models have seen a growing number of applications in the past few years. We will see in the next subsections some of these, and how using this modelling approach gave some insights.

3.1 Variation in survival rates in a sheep population

How density-dependent and density-independent factors affect survival in natural populations has been a long-lasting debate [9]. In a study of Soay sheep (*Ovis aries*), Milner *et al.* showed how a mixed

modelling approach to survival analysis affected their conclusions [28]. They modelled the annual variation in survival rates as a random effect since such variations are more likely to be stochastic than deterministic. They also considered individual as a random effect in the analysis of adult survival, since some individuals contributed several times to the observations, but the resulting variance component was very small.

Milner *et al.* [28] used an approach where they first selected fixed terms with a GLM in order to avoid fitting many mixed models. Then they added random terms, and further tested the fixed terms with Wald statistics. The crucial point is that if the random effects were ignored and considered as deterministic, as often done in survival analyses, this would result in masking other effects on survival, such as those of body weight, sex, or age. The estimated variance components of the year effect on survival were 0.55 (juveniles), 0.89 (adult females), and 1.93 (adult males).

3.2 Geographic variation in life-history traits of a vole

Variation in life-history traits is the corner-stone of most evolutionary theories, but its quantitative assessment remains difficult [23]. Particularly, quantifying the respective influence of genes, environment, and parental effects on phenotypic variation remains a critical issue. Hansen & Boonstra studied variation in some life-history traits of the meadow vole (*Microtus pennsylvanicus*) [12]. They used mixed models in order to quantify the effects of environment, maternal influence, genetic differentiation among population, and additive genetic variance within populations, all considered as potential random effects. They used an AIC-based approach (see Appendix 1) to select the significant random effects.

In this study, the interest was clearly in estimating the variance components rather than taking them into account in order to model other (fixed) effects. The major results were that geographic variation in body size and growth could not be accounted for by genetic differences, whereas within population variation was dominated by additive genetic and maternal effects, both combined explaining about 40 % of the variance.

3.3 Modelling fluctuating asymmetry

A particularly interesting example of variation as a quantity of interest is the case of fluctuating asymmetry. Variations occurring among both sides of bilaterally symmetrical organisms are supposed to reflect the developmental instability of the individuals [24]. Estimation of fluctuating asymmetry may be strongly affected both by measurement error and directional asymmetry.

Van Dongen *et al.* [49] proposed a methodology with which to model FA, directional asymmetry and measurement error using a mixed model. This approach allows one to test for the significance of fluctuating asymmetry, to model and test for heterogeneity in both fluctuating asymmetry and measurement error among samples, to test for non-zero directional asymmetry, to obtain unbiased estimates of individual fluctuating asymmetry levels, and to analyze different traits simultaneously. In this model, fluctuating asymmetry is not estimated as a variation around a mean, as in the two-way mixed-effects analysis of variance, but as a variation of a slope (Fig. 1).

Thomson [48] emphasized that using mixed models in the study of asymmetry may also be applied to characters other than bilateral morphologies, and could be a general approach to quantify developmental stability.

3.4 Modelling Taylor's power law in insects

Taylor's power law is probably one of the most famous ecological theory where variability is the focus of interest [45]. In simple words, it states that the variability in abundance of species increases with their mean abundances. Recently, Candy compared several approaches to the analysis of insect count data from multi-level sampling [5]. Mixed models and traditional approaches based on the use of different kinds of analysis of variance generally gave similar results. However, the latter approaches sometimes estimated negative variance components, a side-effect not encountered with mixed models which makes them more appropriate for prediction. Furthermore, mixed models can accommodate the non-normality of observations (which is obvious with count data) through using a generalized modelling framework.

Candy built on previous works [4] to develop an approach that combines the features of mixed models with those of GENERALIZED ADDITIVE MODELS to include a non-parametric function of the predictors [13,55]. This is an interesting perspective, but which should

be considered with caution since it is likely to combine the difficulties in fitting generalized mixed models (see Appendix 2) with those of choosing a smooth function for the non-parametric terms [50,52].

3.5 Variation in substitution rates in DNA and phylogenetic inference

To take into account the variation in substitution rate along a DNA sequence when reconstructing a phylogeny, Yang developed a method where rates at different sites are taken to be random variables from a gamma distribution [56]. He used a mixed model approach to perform estimation with maximum likelihood [59]. However, in its original form, this approach was only tractable for the analysis of very small samples (6 or 7 species). Yang further proposed to discretize the gamma distribution so that substitution rates vary through discrete categories [57], which permits the analysis of large data sets [58].

This mixed modelling approach to phylogeny reconstruction considerably improves the fit of molecular evolution models to phylogenetic data, and has been used to demonstrate the inadequacy of other simpler methods of phylogeny reconstruction [43,53].

4 Correlated data and generalized estimating equations

An important class of methods is provided by generalized estimating equations (GEEs) which can be viewed as extensions of the GLMs in that the different observations are not assumed to be independent. The GEE approach is related to the mixed-effects modelling, but it is easier from a computational point of view [14]. There is however an important distinction between the two approaches: with GEEs, the focus is on the mean structure in the data, and the dependence among observations is taken into account in order to have more efficient parameter estimates and less biased tests [32].

GEEs mixes the traditional formulation of a GLM with a correlation matrix giving the strength of the dependence among observations [20] (Appendix 1). The structure of this matrix can either be estimated from the data, or fixed by the user. The latter situation is of particular interest if the dependence among observations is known a priori, for instance, using genetic relatedness data. As with mixed models, it is typical that observations are made on clusters which are assumed to be independent.

An attractive feature of GEEs is that the analysis of non-normal responses does not raise any special difficulty, whereas this requires intense computations with approximate methods in the case of mixed models (see Appendix 2).

An important property of GEEs is that they are robust to misspecification of the correlation structure, meaning that the GLM analysis associated with GEE fitting is not biased if the user is mistaken in his assumptions about the dependence among observations [20], but the estimation may become less efficient (i.e. the parameter estimates will be less likely to be close to the true values of the parameters) in these situations, particularly if the number of clusters is small [14].

GEEs are widely used in biomedical studies to analyse data collected repeatedly on the same subjects, for instance doses measured on the same patients during a treatment. This easily extends to situations such as the analysis of count data from ecological census where there is a likely serial correlation through time (year-to-year) [46]. In these cases, the generalized modelling framework takes into account the non-normality of the data (which are of course integers) using a Poisson distribution.

5 Graphical methods

There has been tremendous recent progress in the development of graphical methods due mainly to the increase in computer power. These methods give fresh perspectives on exploratory data analyses. The most spectacular developments are those of dynamic graphics: this can be done for a variety of methods (e.g., correlations, multivariate analyses). If a data set has a large number of variables, a bivariate plot can be drawn with the variables on both axes changed dynamically allowing the user to visualize all plots in turn. A more complicated method is the grand tour where all variables (or a subset selected by the user) are projected on a plane, and the contribution of each variable to both axes is changed dynamically, mostly randomly [44].

The increase in computing power has resulted in the concomitant generalization of user-friendly tools (mainly using the computer mouse) for the interactive visualisation of data such as brushing and identification of observations, speed control of grand tour, selection of variables, and so on [44].

Some graphical methods are particularly useful in conjunction with

particular analytical methods. For instance, multi-panel and conditional plots or histograms, where a set of plots or histograms are drawn with respect to one or several categorical variables, is a useful preliminary to mixed modelling [36,37,51] (Fig. 1).

Graphical methods are in a state of continuous development, and the near general availability of more powerful computers will certainly lead to new methods. It is interesting that some software packages already have a programmable graphic environment (e.g., R, S-PLUS; see also Appendix 2) giving the opportunity to the user of creating his own types of graphics.

6 Conclusions and future prospects

The recent developments and enhancements in methods for statistical analysis give ecologists and evolutionists the opportunity to consider variability with a new perspective, and to link tightly theoretical and conceptual models [10,15,31] on one side, and empirical and statistical models on the other side. Variability needs no longer to be considered as a nuisance in data analysis, but can itself be the subject of study. This is an invitation to ecologists and evolutionists to give more emphasis on parameter estimation than is currently done, a progression that has been called for by statisticians for some time [30].

The currently increasing interest in mixed models will surely continue in the future. Since mixed models have been shown to perform as least as well as specialized methods in several fields, they constitute a general and flexible approach to data analysis. Mixed models will be very useful in disciplines, such as genomics [41], where data are accumulating at a fast rate. The possibility of considering some effects as random avoids the difficulties of over-parametrization which plague classical models.

Acknowledgements. We are grateful to Bill Venables, Alastair Grant, Jo Ridley, and two anonymous referees for helpful comments on a previous version of our manuscript.

REFERENCES

- [1] Aitkin M. 1987. Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics. Journal of the Royal Statistical Society. Series C* **36**: 332–339.

- [2] Bjørnstad O.N., Fromentin J.-M., Stenseth N.C. & Gjøsæter J. 1999. A new test for density-dependent survival: the case of coastal cod populations. *Ecology* **80**: 1278–1288.
- [3] Breslow N.E. & Clayton D.G. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**: 9–25.
- [4] Candy S.G. 1997. Estimation in forest yield models using composite link functions with random effects. *Biometrics* **53**: 146–160.
- [5] Candy S.G. 2000. The application of generalized linear mixed models to multi-level sampling for insect population monitoring. *Environmental and Ecological Statistics* **7**: 217–238.
- [6] Clutton-Brock T.H., Brotherton P.N.M., Russell A.F., O’Riain M.J., Gaynor D., Kansky R., Griffin A., Manser M., Sharpe L., McIlrath G.M., Small T., Moss A. & Monfort S. 2001. Cooperation, control, and concession in meerkat groups. *Science* **291**: 478–181.
- [7] Cotgreave P. 1993. The relationship between body size and population abundance in animals. *Trends in Ecology & Evolution* **8**: 244–248.
- [8] Damuth J. 1991. Of size and abundance. *Nature* **351**: 268–269.
- [9] Dennis B., Kemp W.P. & Taper M.L. 1998. Joint density dependence. *Ecology* **79**: 426–441.
- [10] García-Charton J.A. & Pérez-Ruzafa Á. 1999. Ecological heterogeneity and the evaluation of the effects of marine reserves. *Fisheries Research* **42**: 1–20.
- [11] Gregory R.D. & Gaston K.J. 2000. Explanations of commonness and rarity in British breeding birds: separating resource use and resource availability. *Oikos* **88**: 515–526.
- [12] Hansen T.F. & Boonstra R. 2000. The best in all possible worlds? A quantitative genetic study of geographic variation in the meadow vole, *Microtus pennsylvanicus*. *Oikos* **89**: 81–94.
- [13] Hastie T.J. & Tibshirani R.J. 1990. Generalized additive models. Chapman & Hall, London.
- [14] Horton N.J. & Lipsitz S.R. 1999. Review of software to fit generalized estimating equation regression models. *American Statistician* **53**: 160–169.
- [15] Kelly E.J. & Campbell K. 2000. Separating variability and uncertainty in environmental risk assessment — Making choices. *Human and Ecological Risk Assessment* **6**: 1–13.
- [16] Kleinschmidt I., Sharp B.L., Clarke G.P.Y., Curtis B. & Fraser C. 2001. Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in KwaZulu Natal, South Africa. *American Journal of Epidemiology* **153**: 1213–1221.
- [17] Krackow S. & Tkadlec E. 2001. Analysis of brood sex ratios: implications of offspring clustering. *Behavioral Ecology and Sociobiology* **50**:

293–301.

- [18] Lee Y. & Nelder J.A. 2000. The relationship between double-exponential families and extended quasi-likelihood families, with application to modelling Geissler’s human sex ratio data. *Applied Statistics. Journal of the Royal Statistical Society. Series C* **49**: 413–419.
- [19] Lee Y. & Nelder J.A. 2000. Two ways of modelling overdispersion in non-normal data. *Applied Statistics. Journal of the Royal Statistical Society. Series C* **49**: 591–598.
- [20] Liang K.-Y. & Zeger S.L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13–22.
- [21] Lin X. 1997. Variance component testing in generalised linear models with random effects. *Biometrika* **84**: 309–326.
- [22] Littell R.C., Pendergast J. & Natarajan R. 2000. Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine* **19**: 1793–1819.
- [23] Lynch C.B. & Walsh B. 1998. Genetics and analysis of quantitative traits. Sinauer, Sunderland, Mass., USA.
- [24] Markow T.A. 1995. Evolutionary ecology and developmental instability. *Annual Review of Entomology* **40**: 105–120.
- [25] Merilä J., Kruuk L.E.B. & Sheldon B.C. 2001. Cryptic evolution in a wild bird population. *Nature* **412**: 76–79.
- [26] Merilä J., Kruuk L.E.B. & Sheldon B.C. 2001. Natural selection on the genetical component of variance in body condition in a wild bird population. *Journal of Evolutionary Biology* **14**: 918–929.
- [27] Mikulich S.K., Zerbe G.O., Jones R.H. & Crowley T.J. 1999. Relating the classical covariance adjustment techniques of multivariate growth curve models to modern univariate mixed effects models. *Biometrics* **55**: 957–964.
- [28] Milner J.M., Elston D.A. & Albon S.D. 1999. Estimating the contributions of population density and climatic fluctuations to interannual variation in survival of Soay sheep. *Journal of Animal Ecology* **68**: 1235–1247.
- [29] Morrell C.H. 1998. Likelihood ratio testing of variance components in the linear mixed effects model using restricted maximum likelihood. *Biometrics* **54**: 1560–1568.
- [30] Nelder J.A. 1999. From statistics to statistical science (with discussion). *Statistician. Journal of the Royal Statistical Society. Series D* **48**: 257–269.
- [31] Palmer M.A., Hakenkamp C.C. & Nelsonbaker K. 1997. Ecological heterogeneity in streams: why variance matters. *Journal of the North American Benthological Society* **16**: 189–202.
- [32] Palmgren J. 2000. Exponential family models and statistical genetics. *Statistical Methods in Medical Research* **9**: 57–72.

- [33] Paradis E., Baillie S.R., Sutherland W.J. & Gregory R.D. 2002. Exploring density-dependent relationships in demographic parameters in natural populations of birds at a large spatial scale. *Oikos* **97**: 293–307.
- [34] Patterson H.D. & Thompson R. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**: 545–554.
- [35] Paul R.E.L., Coulson T.N., Raibaud A. & Brey P.T. 2000. Sex determination in malaria parasites. *Science* **287**: 128–131.
- [36] Pinheiro J.C. & Bates D.M. 1995. Model building for nonlinear mixed-effects models. Technical Report No.91, Department of Biostatistics, University of Wisconsin, Madison.
- [37] Pinheiro J.C. & Bates D.M. 1997. Graphical methods for data with multiple levels of nesting. 1997 Joint Statistical Meetings,
- [38] Platt R.W., Leroux B.G. & Breslow N. 1999. Generalized linear mixed models for meta-analysis. *Statistics in Medicine* **18**: 643–654.
- [39] Preston B.T., Stevenson I.R., Pemberton J.M. & Wilson K. 2001. Dominant rams lose out by sperm depletion. *Nature* **409**: 681–682.
- [40] Rodríguez G. & Goldman N. 2001. Improved estimation procedures for multilevel models with binary response: a case-study. *Journal of the Royal Statistical Society. Series A. Statistics in Society* **164**: 339–355.
- [41] Rodríguez-Zas S.L. & Southey B.R. 2001. Analysis of nucleotide sequence data using mixed model methodology. *Genetic Epidemiology* **21**: S638–S642.
- [42] Serrano D., Tella J.L., Forero M.G. & Donazar J.A. 2001. Factors affecting breeding dispersal in the facultatively colonial lesser kestrel: individual experience vs. conspecific cues. *Journal of Animal Ecology* **70**: 568–578.
- [43] Sullivan J. & Swofford D.L. 1997. Are Guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *Journal of Mammalian Evolution* **4**: 77–86.
- [44] Swayne D.F., Cook D. & Buja A. 1998. XGobi: interactive dynamic data visualization in the X window system. *Journal of Computational and Graphical Statistics* **7**: 113–130.
- [45] Taylor L.R. 1961. Aggregation, variance and the mean. *Nature* **189**: 732–735.
- [46] Thomas L. 1996. Monitoring long-term population change: why are there so many analysis methods? *Ecology* **77**: 49–58.
- [47] Thompson S.G. & Sharp S.J. 1999. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* **18**: 2693–2708.
- [48] Thomson D.L. 1999. Intraindividual variance in trait size and the analysis of developmental instability. *Animal Behaviour* **57**: 731–734.

- [49] Van Dongen S., Molenberghs G. & Matthysen E. 1999. The statistical analysis of fluctuating asymmetry: REML estimation of a mixed regression model. *Journal of Evolutionary Biology* **12**: 94–102.
- [50] Venables W.N. 2000. Exegeses on linear models. Paper presented to the S-PLUS User’s Conference, 8-9th October 1998, Washington, DC.
- [51] Venables W.N. & Ripley B.D. 1999. Modern applied statistics with S-PLUS (third edition). Springer, New York.
- [52] Verbyla A.P., Cullis B.R., Kenward M.G. & Welham S.J. 1999. The analysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics. Journal of the Royal Statistical Society. Series C* **48**: 269–300.
- [53] Whelan S., Liò P. & Goldman N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics* **17**: 262–272.
- [54] Williams J.T. & Blangero J. 1999. Comparison of variance components and sibpair-based approaches to quantitative trait linkage analysis in unselected samples. *Genetic Epidemiology* **16**: 113–134.
- [55] Wood S.N. 2001. Partially specified ecological models. *Ecological Monographs* **71**: 1–25.
- [56] Yang Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10**: 1396–1401.
- [57] Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39**: 306–314.
- [58] Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution* **11**: 367–372.
- [59] Yang Z. & Wang T. 1995. Mixed model analysis of DNA sequence evolution. *Biometrics* **51**: 552–561.

Appendix 1. Modelling dependency among observations

Dependency among observations means that the value of a particular observation is partially influenced by the values of others. This translates into non-null correlations among observations, and can be modelled with a correlation matrix. For a standard linear model, this matrix would have 1’s on its diagonal (each observation is equal to itself) and 0’s elsewhere (all observations are independent). In methods such as mixed models or GEEs, this is relaxed by assuming that the data are structured in clusters which are independent, but observations within a cluster are dependent and their correlation structure is the same. Clusters could be individuals on which a variable is measured through time (temporal correlation), families in which several members are studied (genetic correlation), plots sampled repeatedly (spatial correlation), etc.

Table A1 gives some examples of correlation structure that include those most commonly available in statistical softwares (exchangeable, unstructured, and auto-regressive). Examples of the corresponding correlation matrix with four observations per cluster ($c = 4$) are given. A correlation structure can also be defined using a function of some measure of dissimilarity between observations: three examples are given that use physical distance. In all cases, the values in the correlation matrix can be fixed by the analyst so that no correlation parameter has to be estimated with the data (in the case of GEEs, this is called a fixed correlation structure). Note that if the parameters of the correlation matrix are fixed, then a single cluster may be analysed and the regression parameters and their standard-errors estimated taking the dependency among observations into account.

Table A1. Examples of correlation structure (c : cluster size)

Structure	example				number of parameters
Independence	1	0	0	0	0
	0	1	0	0	
	0	0	1	0	
	0	0	0	1	
Exchangeable (or compound symmetric)	1	ρ	ρ	ρ	1
	ρ	1	ρ	ρ	
	ρ	ρ	1	ρ	
	ρ	ρ	ρ	1	
Unstructured (or general symmetric)	1	$\rho_{1,2}$	$\rho_{1,3}$	$\rho_{1,4}$	$c(c-1)/2$
	$\rho_{1,2}$	1	$\rho_{2,3}$	$\rho_{2,4}$	
	$\rho_{1,3}$	$\rho_{2,3}$	1	$\rho_{3,4}$	
	$\rho_{1,4}$	$\rho_{2,4}$	$\rho_{3,4}$	1	
Auto-regressive	1	ρ	ρ^2	ρ^3	1
	ρ	1	ρ	ρ^2	
	ρ^2	ρ	1	ρ	
	ρ^3	ρ^2	ρ	1	
Exponential spatial correlation	$\rho_{ij} = (1 - N) \exp\left(-\frac{d_{ij}}{R}\right)$				2 (N : nugget effect, R : range)
Gaussian spatial correlation	$\rho_{ij} = (1 - N) \exp\left[-\left(\frac{d_{ij}}{R}\right)^2\right]$				idem
Linear spatial correlation	$\rho_{ij} = (1 - N) \left(1 - \frac{d_{ij}}{R}\right)$				idem

Appendix 2. Mixed-effects models

Mixed-effects models (or mixed models in short) are regression models where some coefficients are constant (like in a standard regression), and others vary randomly. Mixed models are sometimes called *random coefficients models* or *multi-stratum* (or *multi-levels*) models depending whether the emphasis is on random variation in the coefficients or on the structure of the data.

Determining the structure of a mixed model, that is which effects are considered fixed and/or random, is a critical step. In some situations, this can be done using prior information on the data (such as biological information). Otherwise, the usual approach of testing for the significance of random and fixed effects (similar to a stepwise procedure) can be used, but this can result in fitting a very large number of models if there are many potential effects. If this is the case, it is recommended to use a model selection criterion such as the AKAIKE INFORMATION CRITERION (AIC) or the SCHWARZ INFORMATION CRITERION (SIC) [22,36]. These criteria are particularly appropriate to compare models with the same fixed effects. The SIC is more conservative with respect to the number of parameters in the models than the AIC; thus, if the choice of a parsimonious model is the objective, the former should be preferred.

An alternative exists for choosing which effects should be considered as random and which ones as purely fixed: this consists of fitting a model that includes all candidate effects as mixed, then an examination of eigenvalues of the estimated variance-covariance matrix indicates which random effects can be dropped from the model [36].

In many practical situations, such as unequal sample sizes among clusters, the maximum likelihood estimators are biased because of the estimation of variance components, and RESTRICTED MAXIMUM LIKELIHOOD (REML) should be used [34]. REML can be used to test if a random effect is significantly different from zero (i.e. its associated variance is non-null). However, in this framework, the distribution of the test statistic (REML ratio test) does not follow the usual χ^2 distribution, but rather a 50:50 mixture of χ^2 distributions [29].

Like the linear model, the linear mixed-effects model (denoted LME) can be generalized in different ways, the most common are the nonlinear mixed-effects model (NLME), and the generalized linear mixed-effects model (GLMM).

Fitting GLMMs is more difficult than fitting LMEs because of the need to integrate the likelihood over the random effects. Several approaches have been proposed which may be classified in two broad categories: approximations of the high-dimensional integration, and Bayesian methods using computer intensive algorithms [32]. Currently, the most “practical” approach seems to be based on penalized quasi-likelihood (PQL) [3]. A wise choice seems to fit GLMMs with two methods (PQL and Bayesian) and compare the consistency of the results [40].

The significance of fixed effects in a GLMM can be tested with the usual Wald test which follow a χ^2 distribution. Lin [21] proposes a global test for the null hypothesis that all variance components in a GLMM are zero, and tests of the individual variance components separately; this procedure does not require fitting of the mixed model, and is robust to misspecification of the joint distribution of the variance components.

Appendix 3. Softwares and packages

General softwares

- Genstat® performs analysis of LMEs (possibly multivariate), of repeated measurements, spatial models, and GLMMs. There are also extensive graphics facilities. Genstat is available for a wide range of systems. <http://www.vsn-intl.com/genstat/index.htm>
- LISREL is particularly oriented for the analysis of structural equations models, and includes LMEs and NLMEs for normal data. It is available for Windows, Macintosh and Unix/Linux. <http://www.ssicentral.com/lisrel/mainlis.htm>
- LispStat is written in the Lisp language. It has an extensive graphic environment, including dynamic and spinning plots. Binaries for Windows, Macintosh, and some Unix systems, as well as the source code are available at <http://www.stat.umn.edu/~luke/xls/xlsinfo/xlsinfo.html>.
- R is a language defined as a dialect of S (see S-PLUS below). Most of the facilities of interest here are included in packages that must be installed separately from the base system of R: they allow analysis with LMEs, NLMEs, GLMMs, and GEEs. Some packages specialize in the analysis of spatial data. A module that links ggobi with R is distributed at the ggobi Web-site (see below). Binaries for Windows, Macintosh, and some Unix/Linux systems, as well as the source code are available at <http://cran.r-project.org/>. The package nlme has its own Web-page with documentation, and is similar to the one in S-PLUS <http://nlme.stat.wisc.edu/>.
- SAS® contains procedures to fit LMEs, NLMEs, and GEEs. Macros, particularly for GLMMs, are freely distributed at <http://ewe3.sas.com/techsup/download/stat/>. SAS is available for most operating systems, general information can be found at <http://www.sas.com/>.
- S-PLUS™ is the commercial implementation of the S language. It includes functions to fit LMEs and NLMEs. Many programs in S are freely available (e.g., on statlib at <http://stat.cmu.edu/S/>). S-PLUS is available for Windows and Unix/Linux. <http://www.insightful.com/>

- Stata® contains procedures to fit LMEs, GEEs, generalized least squares, and GLMMs for count data, and a number of corrections to correct for variance heterogeneity. Stata is available for Windows, Macintosh and Unix. <http://www.stata.com/>
- ViSta is a software written in Lisp which is based on Lisp-Stat. It includes flexible dynamic graphics, and a module for multi-level modelling. Binaries for Windows and Macintosh, and the Lisp codes for Unix are freely available at <http://forrest.psych.unc.edu/research/>.

Specialized softwares

- aML is a software for multilevel models. It can analyse continuous, binomial, and negative binomial count responses. Binaries for Windows and Unix/Linux are available at <http://www.applied-ml.com/>.
- BUGS is a program that carries out Bayesian analysis of complex models which are specified with a declarative language through a graphical interface. The program is distributed in different forms, see <http://www.mrc-bsu.cam.ac.uk/bugs/>.
- DFREML is a suite of programs to estimate (co)variance components or covariance functions, and the resulting genetic parameters by REML fitting of an animal model. DFREML is available free of charge to the scientific community. The code in Fortran 90, and binaries for Unix/Linux and DOS are available at <http://agbu.une.edu.au/~kmeyer/dfreml.html>.
- Egret is a Windows program that includes notably the GLMM for binomial responses (called logistic regression with random effects in this program). Information available at <http://www.cytel.com/new.pages/EGRET.2.html>.
- ggobi is a software for dynamic graphics with several facilities. It is currently in development and will supersede XGobi. Binaries for Windows and the codes are available at <http://www.ggobi.org/>.
- HLM is a Windows program that fits several linear and non-linear models with random effects for normal and non-normal responses. Information available at <http://www.ssicentral.com/hlm/hlm.htm>.
- MANET is a software for data visualization and exploration which is remarkable for its interactivity. It is mentioned here because it includes conditional graphics though in a restricted implementation compared to what is available in R or S-PLUS. Binaries for Macintosh are freely distributed at <http://www1.Math.Uni-Augsburg.DE/Manet/>.
- MLwiN is a software package for fitting multilevel models. Binaries for Windows are available: <http://multilevel.ioe.ac.uk/index.html>.

- Orca is a program for interactive and dynamic graphics. Binaries (requiring Java) and codes are available at <http://software.biostat.washington.edu/statsoft/orca>.
- PEST is a Fortran program for multivariate prediction and estimation with breeding data; it covers fixed, random and mixed models. Further informations are available at <ftp://ftp.tzv.fal.de/pub/pest/doc/>.
- SUDDAN is a software specialized in the analysis of correlated data with GEEs; it implements many variants of this method. It is available for Windows and Solaris at <http://www.rti.org/sudaan/>.
- TRIM is a free Windows software for the analysis of population trends data. It fits a log-linear model to population counts with GEEs. It is available from Statistics Netherlands, PO Box 4000, 2270 JM Voorburg, Netherlands.
- VCE is a program to estimate covariance matrices in a rather general manner, it is oriented to the analysis of breeding data. It is freely available for research purposes. Binaries for Windows and Unix/Linux are available at <http://www.tzv.fal.de/institut/genetik/vce4/vce4.html>.
- XGobi and XGvis are two programs written in C for high-dimensional data visualization. XGobi has many tools including dynamic graphics. XGvis performs several forms of multidimensional scaling, and uses XGobi as its visualization engine. Codes are freely available at <http://www.research.att.com/areas/stat/xgobi/>.

Appendix 4. Glossary

Akaike information criterion (AIC): a likelihood-based model selection criterion which can compare nested or non-nested models. The AIC is a trade-off between the fit of the model to the data (measured by its likelihood) and the number of parameters in the model (models with more parameters will tend to fit data better).

Best linear unbiased predictor (BLUP): the best predictor, for any type of model, requires that the distribution of the random variables be known. The best predictor is the conditional mean of the predictor given the data vector, which is unbiased and has the smallest mean squared error of all predictors. However, in general the required distributions are not known, though a linear function can be used instead.

Generalized additive models (GAMs): this is an extension of the GLMs where the relationship between the predictors and the response is modelled in a non-parametric way using smooth (continuous) functions.

Generalized linear models (GLMs): this is an approach for the linear regression of a wide range of data (continuous, discrete, counts). Its two main features are: (a) a linear model of the transformed mean $g(\mu) = \beta X$,

where g is called a link function (note that the mean is transformed, not the observations), (b) the dispersion of the data around the expected mean is given by $\text{Var}(y) = \phi V(\mu)$ with ϕ the dispersion parameter, and $V(\mu)$ the variance function which is given with respect to the assumed distribution of y (ϕ is fixed or estimated from the data depending on this distribution). The GLM includes several methods (standard linear regression, logistic regression, log-linear Poisson regression, analysis of variance with fixed effects) as special cases.

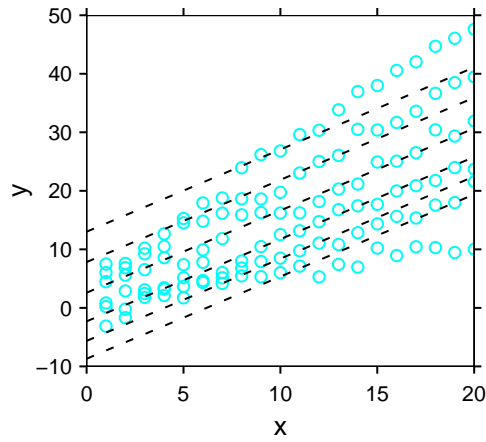
Overdispersion: this concept applies only when there is an expected variance of the observations. This is the case mainly for count data, for instance, if a Poisson distribution is assumed, then the variance is expected to be equal to the mean $\text{Var}(y) = E(y)$; if a binomial distribution is assumed then we expect $\text{Var}(y) = E(y)[1 - E(y)]$.

Restricted (residual) maximum likelihood (REML): a modification of the maximum likelihood (ML) estimation method when some effects in the model are random, thus avoiding some biases in the ML estimates (though the word *restricted* is commonly used, *residual* is the correct one).

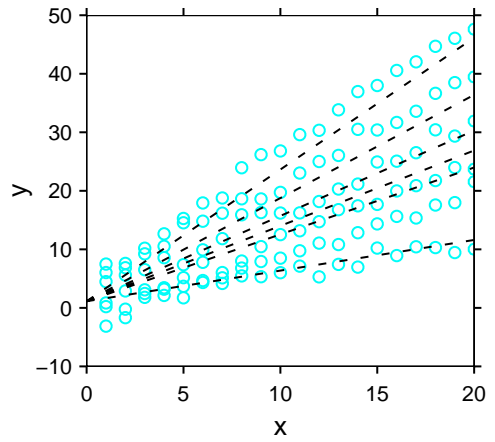
Schwarz information criterion (SIC): this is also called the Bayesian information criterion (BIC, or sometimes simply SC or BC). It is similar to the AIC, but in addition to the number of parameters, a penalty is given with respect to (the log-transformed) sample size. The SIC will tend to select more parsimonious models than the AIC.

Fig. 1. Illustration of the mixed-effects modelling approach. Suppose two variables (denoted x and y) are observed 20 times on 6 clusters. The interest is in modelling a linear relationship between x and y but this is likely to vary among clusters. Thus we may assume that the parameters of the linear model (the intercept and the slope, denoted α and β , respectively) are random variables following a normal distribution each characterized by a mean (μ_α and μ_β) and a standard-deviation (σ_α and σ_β). (a) This random intercept model assumes a fixed effect of x on y ($\sigma_\beta = 0$) and a random intercept per cluster ($\alpha \sim N(\mu_\alpha, \sigma_\alpha^2)$). Considering the intercept as random implies the estimation of two parameters (μ_α and σ_α) whatever the number of clusters, whereas fitting different lines to each cluster would imply the estimation of as many intercepts as clusters. It is interesting to note that if we assume a null slope ($\beta = 0$ for all clusters) then the resulting random intercept model is similar to a one-way analysis of variance with a single random factor; thus the model represented on the figure can be viewed as an analysis of covariance with a fixed continuous covariate and a single random factor. Note that estimating the parameters of the distribution of the intercepts (the variance component), and estimating the actual values of these intercepts for each cluster (the random effects) are two distinct issues; the dashed lines with the different intercepts are here to illustrate the idea of a random intercept. (b) The random slope model represented here assumes $\beta \sim N(\mu_\beta, \sigma_\beta^2)$, and $\sigma_\alpha = 0$. It is possible to assume $\mu_\beta = 0$, $\sigma_\beta \neq 0$, that is no fixed effect of x on y , only a random effect of cluster (see the application to the analysis of fluctuating asymmetry). (c) Scatterplot of y vs. x conditioned on cluster (labeled A through F): this suggests that a model with both random intercept and slope might fit these data (they were actually simulated with $y_{ij} = \beta_j x_{ij} + \alpha_j + \epsilon_{ij}$, where $\beta_j \sim N(1, 1)$, $\alpha_j \sim N(1, 4)$, $\epsilon_{ij} \sim N(0, 1)$, $i = 1, \dots, 20$, and $j = 1, \dots, 6$).

(a) Random intercept model



(b) Random slope model



(c)

