Analysis of comparative data using generalized estimating equations

Julien Claude, Emmanuel Paradis

Analysis of comparative data using generalized estimating equations

EMMANUEL PARADIS & JULIEN CLAUDE

Laboratoire de Paléontologie, Paléobiologie & Phylogénie,

Institut des Sciences de l'Évolution,

Université Montpellier II, F-34095 Montpellier cédex 05, France

Running headline: comparative data and GEE

Correspondence:
Emmanuel Paradis
Laboratoire de Paléontologie, Paléobiologie & Phylogénie
Institut des Sciences de l'Évolution
Université Montpellier II
F-34095 Montpellier cédex 05
France

phone: +33 4 67 14 39 64
fax: +33 4 67 14 36 10
e-mail: paradis@isem.univ-montp2.fr

Abstract

It is widely acknowledged that the analysis of comparative data from related species should be performed taking into account their phylogenetic relationships. We introduce a new method, based on the use of generalized estimating equations, for the analysis of comparative data. The principle is to incorporate, in the modelling process, a correlation matrix that specifies the dependence among observations. This matrix is obtained from the phylogenetic tree of the studied species. Using this approach, a variety of distributions (discrete or continuous) can be analysed using a generalized linear modelling framework, phylogenies with multichotomies can be analysed, and there is no need to estimate ancestral character state. A simulation study showed that the proposed approach has good statistical properties with a type I error rate close to the nominal 5%, and statistical power to detect correlated evolution between two characters which increases with the strength of the correlation. The proposed approach performs well for the analysis of discrete characters. We illustrate our approach with some data on macro-ecological correlates in birds. Some extensions of the use of generalized estimating equations are discussed.

1. Introduction

Comparing species is central in many biological issues. For instance, one of Darwin's (1859) main point in his theory of evolution by natural selection was to show that "wide-ranging, much diffused, and common species vary most." More recently, scaling and allometric relationships have been widely used in biology and physiology (Calder, 1983; Peters, 1983; Schmidt-Nielsen, 1984). Traditionally, these relationships were studied using statistical methods (mainly regressions) which assume that species are independent observations. However, species are not independent observations because they are linked by their phylogenetic relationships. Several methods have been proposed to analyse interspecific comparative data taking this difficulty into account (Cheverud *et al.*, 1985; Felsenstein, 1985; Grafen, 1989; Gittleman & Kot, 1990; Lynch, 1991; among others).

The method of phylogenetically independent contrasts (referred to as the contrasts method in this paper) substitutes the original data collected on species by a set of contrasts computed between pairs of sister-species, and between pairs of nodes in the reconstructed phylogeny (Felsenstein, 1985). The character states at the node are reconstructed under a model of random evolution (Felsenstein, 1985). This method is certainly the most widely used to deal with the problem of non-independence among species. Recently, Martins & Hansen (1997) presented a method which can be viewed as an extension of Grafen's (1989) method, based on the analysis of a linear model with generalized least squares (GLS) where the

variances and covariances of the data are defined with respect to the phylogenetic relationships among species.

In this paper, we present a new approach for the analysis of comparative data taking into account the phylogenetic relationships among species. The traits analysed can be either continuous or discrete, and more than two traits can be analysed simultaneously in a generalized linear modelling framework. The model is fitted using generalized estimating equations where the dependence among species is taken into account with a correlation matrix. We run simulations to assess some statistical properties of our approach. We illustrate its use with an analysis of some data on birds (Paradis *et al.*, 1999).

2. A generalized estimating equations approach

Felsenstein (1985) presented the statistical problem with comparative data when closely related species are included in a sample. These species should, in fact, be considered as pseudo-replications of the same observations, and considering them as independent inflates the number of degrees of freedom in the analysis, resulting in an increased type I error rate. For instance, if in a sample of eight species there are four pairs of sibling-species, then the number of degrees of freedom in a linear regression model should be two, and not six (the critical values for a t-test with a significance level of 5% are 2.45 for df = 6, and 4.30 for df = 2).

The strength of the dependence among the observations on two species depends on the phylogenetic distance between them, and how the observed

characters evolve through time. Thus, a measure of this dependence can be derived from the distances measured on a phylogenetic tree. An appropriate transformation of these distances, depending on the mode of character evolution, gives a matrix of correlations among species: this matrix is symmetric and its diagonal elements equal to unity (Fig. 1). Generalized estimating equations (GEE) are a procedure to fit regression models taking the correlations among the observations into account (Liang & Zeger, 1986). The correlation matrix can take different structures, and its parameter(s) can either be estimated from the data, or fully specified; in the latter case, the structure of the correlation matrix is said to be fixed. In all cases, the parameters from the regression model are estimated from the data. This regression model is a generalized linear model (GLM, McCullagh & Nelder, 1989), meaning that the response variable may be non-normal, for instance, binomial or Poisson.

Consider that we have n species. The phylogenetic relationships among the species are known (topology and branch lengths), and let D be the $n \times n$ matrix of pairwise phylogenetic distances. Let $\boldsymbol{y}$ be a $n \times 1$ vector of a variable (called the response) whose elements are denoted y_i (where the subscript i denotes the ith species, $i = 1, \ldots, n$), and X be the $n \times p$ matrix of covariates (or predictors). Let us assume that the y_i follow a marginal distribution belonging to the exponential distribution family, making possible a generalized linear regression of $\boldsymbol{y}$ with respect to X:

$$g(E[y_i]) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters, $\boldsymbol{x}_i^{\mathrm{T}}$ is the transposed vector of the covariate values for the ith species, and g is a link function. Equation (1) is identical to $g(E[y_i]) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$, where β_0 is an intercept. This does not differ from a standard GLM. Recall that a GLM is a linear model of the expected mean of a variable belonging to the exponential distribution family (e.g., Gaussian, gamma, Poisson, binomial, for the most commonly used distributions), the expected mean being transformed using a link function (McCullagh & Nelder, 1989). In a GLM, the variance of the responses is given by:

$$Var(y_i) = \phi \mathcal{V}(E[y_i]),$$

where ϕ is the dispersion parameter, and $\mathcal{V}(E[y_i])$ is the variance function, both of them being defined with respect to the distribution assumed for y. Note that this is the variance expected under the assumption that all species are independent observations, which we do not want to assume here. If the observations are not independent, we can define the variance-covariance matrix among observations with (Liang & Zeger, 1986):

$$V = \phi A^{1/2} R A^{1/2}, \tag{2}$$

where A is a $n \times n$ diagonal matrix defined by $\mathrm{diag}\{\phi \mathcal{V}(E[y_i])\}$, that is a matrix with all its elements null except the diagonal which contains the variances of the n observations expected under the marginal model, and R is the correlation matrix of the elements of y. If the observations are in fact independent, then R is

an $n \times n$ identity matrix.

3. Estimating parameters and testing hypotheses

The main objective of the present approach is to look for relationships between $\boldsymbol{y}$ and X. An advantage of the GLM setting is that the y_i can follow several types of distributions including Gaussian, gamma, Poisson, or binomial. The two latter distributions are appropriate to model, for instance, numbers and frequencies, respectively. The variables included in X could be continuous or categorical, and the model can include additive, interactive, and nested effects among these predictors.

From the assumptions in the previous section, it is possible to define estimating equations which are consistent estimators of the regression parameters $\boldsymbol{\beta}$. These generalized estimating equations are:

$$\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}} V^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) = 0, \tag{3}$$

where $\boldsymbol{\mu}$ is the $n \times 1$ vector of the mean expected responses whose elements μ_i $(= E[y_i])$ are given by $g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, with g^{-1} as the reciprocal of the link function in equation (1). The Appendix describes a procedure to solve equation (3), and several estimators of the standard-errors of $\boldsymbol{\beta}$.

The estimates of $\boldsymbol{\beta}$ (denoted $\hat{\boldsymbol{\beta}}$) can be found with GEE using information from D and the phylogenetic tree as the correlation matrix R (Fig. 1). We will assume that no element from the matrix R needs to be estimated, so the

correlation structure is fixed.

GEE can be fitted easily with existing software (Horton & Lipsitz, 1999). The estimates of the standard-errors (SE) of $\hat{\boldsymbol{\beta}}$ can be used to test for the significance of each parameter in the model: under the null hypothesis that $\beta = 0$, the ratio $\hat{\beta}/\mathrm{SE}(\hat{\beta})$ follows a t distribution. In the present implementation of GEE, the number of degrees of freedom of this t-test is not the usual residual number of degrees of freedom of the fitted model (that is the sample size minus the number of estimated parameters): since there is a single cluster of observations, this residual number of degrees of freedom actually overestimates the true number of degrees of freedom. In order to correct for this bias, we used a procedure where the number of degrees of freedom is calculated from the phylogenetic tree. The principle of this procedure is to consider such a tree as a representation of the inter-dependence among observations, and thus as the number of degrees of freedom in the data. First consider, for simplicity, that the tree is ultrametric, each tip being equally distant from the root. The number of degrees of freedom is counted by adding, at each dichotomous node in the tree, one unity multiplied by the ratio of the distance from the node to the tips on the distance from the root to the tips. If the node is trichotomous, then two units (multiplied by the same ratio) are added, if the node is tetrachotomous, three units are added, and so on. This is done starting from the root where two degrees of freedom are counted. We denote the final number df_P. This procedure can be

generalized to a non-ultrametric tree with:

$$df_P = \frac{\sum_{\text{tree}} \text{branch length}}{\sum_{i=1}^{n} \text{distance from root to tip}_i} \times n.$$

We will further discuss the need for this correction in the discussion.

To test for the significance of the effect(s) of one or several predictors, we can use an ANOVA-like analysis. Let W denote the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. Under the null hypothesis of no effect, the quantity $\hat{\boldsymbol{\beta}}^{\mathrm{T}} \hat{W}^{-1} \hat{\boldsymbol{\beta}}$ follows an F distribution with df and $df_P - df - 1$ as numbers of degrees of freedom, where df is the number of degrees of freedom associated with $\boldsymbol{\beta}$ (i.e. the number of estimated parameters for the tested effects).

4. Simulation study

In order to assess some properties of the approach we present here, we conducted simulations of character evolution under different scenarios. We considered three phylogenetic trees (Fig. 2). All three trees have 32 tips and the same time of evolution from the root to the tips (155 time-steps). Tree A is a balanced phylogeny with equal branch lengths, tree B has strongly clustered tips, and tree C has some imbalance in its terminal branches. We simulated the evolution of four continuous characters (denote them z_1, z_2, z_3, and z_4) with a Brownian motion model, and eight discrete binary characters taking the value 0 or 1 (y_1 to y_8) with a Markovian model. All characters were set equal to zero at the root of the tree. The continuous characters evolved along the branches by adding, at

each time-step, a random normal variate. After reaching a node, the characters evolved independently on each daughter-branch. The characters z_1 and z_2 evolved in a similar way and independently according to: $z_{1t+1} = z_{1t} + \varepsilon$, where t is the time-step along the branches, and $\varepsilon \sim N(0,1)$. On the other hand, z_3 and z_4 evolved in a correlated way according to: $z_{3t+1} = z_{3t} + \varepsilon$, and $z_{4t+1} = z_{4t} + \zeta$, where $\zeta \sim N(\gamma z_{3t}, 1)$. The parameter γ specifies the strength of the "co-evolution" between z_3 and z_4. We repeated the simulations with four different values of γ: 0.001, 0.005, 0.01, 0.015. The discrete characters had fixed transition probabilities given by a symmetric matrix: the probability of remaining in the same state (either 0 or 1) was p, whereas the probability of changing was $1-p$. The characters y_1 and y_2 evolved independently and with the same transition probabilities ($p = 0.6$), and y_3 and y_4 evolved independently too but with $p = 0.8$. This higher probability is more likely to result in spurious patterns of association between the states of y_3 and y_4 than for y_1 and y_2 (Grafen & Ridley, 1997). The four remaining characters evolved in pair: y_5 evolved in the same way than y_3 and y_4, but y_6 had transition probabilities which depended on the state of y_5 (they were also equal to 0.6 and 0.4). The same model of co-evolution was used for y_7 and y_8, but the transition probabilities were 0.8 and 0.2. The simulations were replicated 1000 times for each parameter value.

At the end of the evolution process, the values of the characters at the tips of the tree were analysed with six regressions using GEE: z_1 on z_2, z_4 on z_3, y_1 on y_2, y_3 on y_4, y_6 on y_5, and y_8 on y_7. The regressions with the continuous characters were done assuming a Gaussian distribution for the response and with

an identity link; the regressions for the discrete characters assumed a binomial distribution for the response, and used a logit link. We expected the test on the significance of the slope to reject the null hypothesis in 5% of the replications (the expected type I error rate since the null hypothesis was true) with the first regression, whereas the rejection rate of the null hypothesis with the second regression will depend on the power of the test (since the null hypothesis was false). The correlation matrix R was derived from the expected variances and covariances of the characters z_1, z_2, z_3, and z_4 at the tips of the trees under a Brownian motion model of evolution (Butler *et al.*, 2000; Garland & Ives, 2000).

To compare the GEE approach with a standard contrasts method, we also analysed the simulated data with contrasts. The contrasts were computed following Felsenstein (1985). For the discrete characters, the values 0 and 1 were used untransformed (see Grafen & Ridley, 1996). A linear correlation coefficient was then computed between the 31 pairs of contrasts, and its statistical significance was assessed using a t-test with 29 degrees of freedom.

Finally, we analysed some of the simulated data with a GLM (i.e., assuming that the observations are independent) in order to assess its type I error rate. We computed for each analysis and each set of parameters, the rate of rejection of the hypothesis of a null slope at a nominal level of 5%. This rate is (i) the type I error rate if the characters were uncorrelated, or (ii) the power of the test if they were correlated. Among a sample of size 1000, a rate of success is significantly greater than 0.05 if 62, or more, successes are observed (this is a one-sided test since we are not interested in the cases where the rejection rate is significantly

smaller than 0.05). Thus, the rejection rate was significantly greater than 5% if we observed 62 or more rejections of the null hypothesis. All these simulations were programmed in R (Ihaka & Gentleman, 1996). The phylogenetically independent contrasts were computed using a C code from the package Phylip (Felsenstein, 1993) called from R.

The results are summarized in Tables 1 and 2. For the continuous characters, the performance of GEE and contrasts were very close; however, the type I error rate of GEE was significantly greater than 5% (around 7%), whereas it was not significantly different from 5% for the contrasts (Table 1). With respect to the power of the tests, the performance of GEE was slightly better than contrasts for the smallest values of γ, but this was the opposite for the greatest values of this parameter (Table 1).

For the discrete characters, GEE performed better: the type I error rates were not significantly different from 5%, except for phylogeny C (Table 2). The type I error rates of contrasts were always significantly greater than 5% for these characters. When the discrete characters evolved in association, the performance were not very good for the lowest value of the parameter ($p = 0.6$). For the greatest value of this parameter ($p = 0.8$), the power of the tests were very similar for the different phylogenies, around 37% for GEE, and around 53% for the contrasts (Table 2).

We evaluated the type I error rate of GLM in the case of continuous characters it was 25.4%, 39.5%, and 25.3% for phylogeny A, B, and C, respectively.

5. Application: relationship between dispersal and population synchrony

There has been a large number of recent publications on the relationships between dispersal and population synchrony (e.g., Lande *et al.*, 1999; Kendall *et al.*, 2000). The strength of synchrony in fluctuations of population density is hypothesized to be driven either by environmental factors (such as climate), or by dispersal connecting local populations (Haydon & Steen, 1997). Interspecific comparisons are obviously valuable to test such hypotheses since population parameters (like dispersal distance) vary much more at the interspecific level than at the intraspecific one. Paradis *et al.* (1999) showed a positive correlation between dispersal distance and population synchrony for 53 species of birds in Britain and Ireland. An examination of the data revealed that other variables obviously influenced the relationships, and GLMs were used to find which among some candidate variables significantly affect population synchrony. The model finally selected has the following form:

population synchrony = dispersal*habitat + long-term national trend, (4)

where the term dispersal*habitat means that the effect of dispersal on synchrony depended on the type of habitat (it was stronger for the species nesting in wet habitats), and the long-term national trend is the slope of the linear change of the global population against time between 1962 and 1995 (the most declining species having more synchronized populations). This left open the issue of the eventual influence of phylogenetic relationships on this relationship since such complex models cannot be fitted directly with contrasts.

We re-analysed these results with GEE separately for breeding and natal dispersal (as done in Paradis *et al.*, 1999). We considered the model described by equation (4). The phylogenetic relationships among the 53 species were taken from Sibley & Ahlquist (1990); we computed the correlation matrix R using the $\Delta T_{50}H$ distances given by Sibley & Ahlquist (1990). The number of degrees of freedom of this tree was $df_P = 16$. The results of the F-tests are summarized in Table 3. For both natal and breeding dispersal, the model was strongly significant indicating that the above relationship was not influenced by the phylogenetic relationships among species. The parameter estimates with their standard-errors are given in Table 4. Interestingly, in this analysis the effect of long-term trend appears as just close to significance, whereas it was strongly significant in the GLM without correcting for phylogeny (Paradis *et al.*, 1999). When looking at the data on population trend, it appears that some of the most closely related species share a common trend: for instance, doves (three species) or tits (five species) are increasing, whereas thrushes (three species) and finches (five species) are declining.

6. Discussion

The analysis of comparative data taking phylogeny into account is now widespread among evolutionists and comparative biologists. Our purpose was to present a new approach that tries to be as general as possible since it has some features of the methods currently available. The GEE approach can be viewed as an extension of the GLS one (Martins & Hansen, 1997) with two differences: the

dependence among observations is specified with a correlation matrix in GEE (a variance-covariance matrix in GLS), and non-normal distribution can easily be modelled with GEE due to a generalized linear modelling framework. GEE permit the use of categorical predictors as well as continuous ones, similarly to the phylogenetic regression (Grafen, 1989), or the GLS method (Martins & Hansen, 1997). Another feature that is common to the GEE approach and some extensions of the contrasts method (Grafen, 1989; Martins & Hansen, 1997) is the possibility to fit complex models that include continuous, categorical predictors, and possible interactions between them.

Our simulation study showed that the approach with GEE has good statistical properties. Unsurprisingly, it had much lower type I error rates than GLM. However, the type I error rates of GEE were slightly greater then 5%, and this difference was statistically different. A possible explanation is that this comes from the fact that the number of degrees of freedom we used in these tests was actually an approximation of the true number. Overall, the contrasts performed very well for continuous characters. This confirms several simulation studies (Martins & Garland, 1991; Díaz-Uriarte & Garland, 1996; Harvey & Rambaud, 1998). For discrete characters, the GEE approach performed globally better than the contrasts one. The type I error rates of GEE were kept smaller than 5%, except for phylogeny C. In all situations the contrasts had inflated type I error rates which certainly comes from the invalid assumption of the distributions of the characters, thus the contrasts method is not robust to violation of this assumption. In terms of power of the tests, they performed very poorly when the

transition probabilities were $\{0.6, 0.4\}$, but this may come simply from the fact these probabilities being too close to 0.5, and thus the co-evolution between the two discrete characters was too weak to be detected. When the transition probabilities were $\{0.8, 0.2\}$, the tests performed better, with greater power for the contrasts than for GEE, but this was obviously at the expense of the increased type I error rate when the null hypothesis was true.

Several recent studies have attempted to compare different approaches to the comparative method (Garland & Ives, 2000; Rohlf, 2001). Garland & Ives (2000) showed that the contrasts and GLS methods are identical and expected to give the same results. In theory, the GEE assuming normal errors and the GLS are identical. It may thus be surprising that the contrasts and the GEE yielded different results in our simulations of continuous characters, though the differences were small (Table 1). It should be noted that GEE and GLS are solved with quite different algorithms which may partially explain this discrepancy. Furthermore, the difficulties in estimating the coefficient standard-errors (see below) may be important here as well.

Additionally, it has been demonstrated that different computer programs may give remarkably different results for much simpler problems than the one of comparative analysis such as summary statistics, analysis of variance, or linear regression (McCullough, 1998, 1999). Whether we should expect similar discrepancies for the different comparative methods and their computer implementations is not clear, and obviously needs further study.

We did not simulate character evolution with phylogenies having

multichotomies because the program code to compute contrasts available for our simulations can only deal with dichotomies (programs that can compute contrasts in presence of multichotomies cannot be called from R). Some contrasts methods are known to be valid in the presence of multichotomies (Grafen, 1989; Purvis & Garland, 1993). The procedure to compute the contrasts and the associated number of degrees of freedom differ, however, depending on whether the multichotomies should be treated as true instantaneous radiations or as a series of unresolved dichotomies (Purvis & Garland, 1993). Since, in the GEE approach, we associate some degrees of freedom to each node, the same correction can be applied in the method proposed here, that is only one (fraction of) degree of freedom should be counted for each unresolved multichotomy, rather than the number of daughter-species minus one.

The need to correct the number of degrees of freedom with GEE is not obvious since the dependence among observations is already taken into account with the correlation matrix. However, it is known that GEE estimators of the standard-errors of the regression coefficients (SE(β)) perform poorly when the number of clusters is low (Horton & Lipsitz, 1999; Mancl & DeRouen, 2001). Our approach was thus to correct the number of degrees of freedom in order to keep the type I error rate at an acceptable level. Furthermore, our simulations showed that the power to detect correlated evolution with GEE is close to that of phylogenetically independent contrasts. However, the estimation of coefficient standard-errors with GEE is still under progress (Mancl & DeRouen, 2001), and some attention will be needed in the context of the application of this method to

the analysis of comparative data.

According to our simulations, the GEE approach seems to perform well for the analysis of discrete characters. Other methods exist for similar analyses (Sillén-Tullberg, 1993; Pagel, 1994; Read & Nee, 1995; among others). Grafen & Ridley (1996) assessed the type I error rates of four such methods. To simulate some null data, they used three phylogenies each with 256 tips (Grafen & Ridley, 1996); thus their results cannot be really compared with ours. Clearly, an extensive comparison of the error rates of these methods in different situations (sample sizes, tree balance, pattern of multichotomy) is needed.

The contrasts method can deal only with continuous variables, whereas we have shown that GEE can deal with a variety of distributions because of the GLM setting. This points to the difference in philosophy betwen both approaches. The contrasts method uses a null model of evolution through the phylogenetic tree which is suitable for continuous variables (the Brownian motion model), and then focuses on the contrasts between pairs of taxa and/or nodes which are expected to be normally distributed under the assumed model of evolution. Our GEE approach, focuses on the relationships between the variables observed on recent species, and takes into account their phylogenetic relationships in parameter estimation and statistical testing. We think that both approaches have their own merits. Nevertheless, we believe that the GEE approach is particularly well-suited to characterise relationships between ecological, biological, and physiological variables when the non-independence among taxa (or, more generally, observations) is a potential problem for

statistical inference. In this respect, the regression coefficients estimated by GEE can be interpreted functionally in the same way than the coefficients estimated by a standard linear regression. This is due to the fact that the focus of GEE is on the mean structure in the data (Palmgren, 2000). For instance, it is possible to compare a theoretically expected value of the coefficient with a 95% confidence interval $\hat{\beta} \pm 1.96 \times \mathrm{SE}(\hat{\beta})$.

Like the contrasts method, the GEE approach makes some assumptions on the model of evolution of the characters through the transformation of phylogenetic branch lengths to obtain the correlation matrix. Though we did not consider such a possibility in this paper, we mentioned above that some parameters of the correlation matrix can be estimated from the data. This is possible by assuming a particular structure of this matrix (exchangeable, unstructured, auto-regressive, among others, see Horton & Lipsitz, 1999 for more details on correlation matrix structures). An interesting potentiality here would be to include parameters of the model of character evolution in the correlation matrix, and estimate them from the data (Liang *et al.*, 1992).

An interesting possibility is to apply different transformations to the elements of the matrix D, or to use different measures for D. For instance, instead of using times of divergence, D could be a matrix of genetic or molecular divergence which is justified if it is assumed that the rate of phenotypic evolution is proportional to the rate of molecular evolution. Thus, it is possible to introduce heterogeneous rates of evolution in the analysis.

GEE can be used to analyse comparative data at the intraspecific level as well.

The concern about phylogenetic dependence has traditionally concentrated on the interspecific level, but the problem exists within a species too, and is probably even more important in this case since the shared history among observations (individuals or local populations) is more recent than in the interspecific case.

To conclude, we present in this paper an approach for the analysis of comparative data taking into account phylogeny by the use of generalized estimating equations. The use of a generalized linear modeling framework allows the analysis of a great variety of models. The focus on the relationships among variables observed on recent species makes possible the interpretation of the relations in terms of constraints.

We are grateful to Rob Freckleton, Christophe Thébaud, and two anonymous referees for helpful comments on a previous version of our paper. This is publication 02-031 of the Institut des Sciences de l'Évolution (Unité Mixte de Recherche 5554 du Centre National de la Recherche Scientifique).

REFERENCES

Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988). *The New S Language.* London: Chapman & Hall.

Butler, M. A., Schoener, T. W. & Losos, J. B. (2000). The relationships between sexual size dimorphism and habitat use in Greater Antillean *Anolis lizards*. *Evolution* **54**, 259-272.

Calder, W. A., III (1983). Ecological scaling: mammals and birds. *Annu. Rev. Ecol. Syst.* **14**, 213-230.

Cheverud, J. M., Dow, M. M. & Leutenegger, W. (1985). The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. *Evolution* **39**, 1335-1351.

Darwin, C. (1859). *On the origin of species by means of natural selection.* London: John Murray.

Díaz-Uriarte, R. & Garland, T. (1996). Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian-motion. *Syst. Biol.* **45**, 27-47.

Felsenstein, J. (1985). Phylogenies and the comparative method. *Am. Nat.* **125**, 1-15.

Felsenstein, J. (1993). *Phylip (Phylogeny Inference Package) version 3.5c*. Seattle, USA: http://evolution.genetics.washington.edu/phylip/phylip.html. Department of Genetics, University of Washington.

Garland, T., Jr. & Ives, A. R. (2000). Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *Am. Nat.* **155**, 346-364.

Gittleman, J. L. & Kot, M. (1990). Adaptation: statistics and a null model for estimating phylogenetic effects. *Syst. Zool.* **39**, 227-241.

Grafen, A. (1989). The phylogenetic regression. *Phil. Trans. R. Soc. Lond. B* **326**, 119-156.

Grafen, A. & Ridley, M. (1996). Statistical tests for discrete cross-species data. *J. theor. Biol.* **183**, 255-267.

Grafen, A. & Ridley, M. (1997). A new model for discrete character evolution. *J. theor. Biol.* **184**, 7-14.

Harvey, P. H. & Rambaud, A. (1998). Phylogenetic extinction rates and comparative methodology. *Proc. R. Soc. Lond. B* **265**, 1691-1696.

Haydon, D. & Steen, H. (1997). The effects of large- and small-scale random events on the synchrony of metapopulation dynamics: a theoretical analysis. *Proc. R. Soc. Lond. B* **264**, 1375-1381.

Horton, N. J. & Lipsitz, S. R. (1999). Review of software to fit generalized estimating equation regression models. *Am. Stat.* **53**, 160-169.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *J. Comput. Graph. Statist.* **5**, 299-314.

Kendall, B. E., Bjørnstad, O. N., Bascompte, J., Keitt, T. H. & Fagan, W. F. (2000). Dispersal, environmental correlation, and spatial synchrony in population dynamics. *Am. Nat.* **155**, 628-636.

Lande, R., Engen, S. & Sæther, B.-E. (1999). Spatial scale of population synchrony: environmental correlation versus dispersal and density regulation. *Am. Nat.* **154**, 271-181.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Liang, K.-Y., Zeger, S. L. & Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *J. R. Statist. Soc. B* **54**, 3-40.

Lynch, M. (1991). Methods for the analysis of comparative data in evolutionary biology. *Evolution* **45**, 1065-1080.

Mancl, L. A. & DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics* **57**, 126-134.

Martins, E. P. & Garland, T., Jr. (1991). Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution* **45**, 534-557.

Martins, E. P. & Hansen, T. F. (1997). Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.* **149**, 646-667.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models (second edition).* London: Chapman & Hall.

McCullough, B. D. (1998). Assessing the reliability of statistical software: Part I. *Am. Stat.* **52**, 358-366.

McCullough, B. D. (1999). Assessing the reliability of statistical software: Part II. *Am. Stat.* **53**, 149-159.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B* **255**, 37-445.

Palmgren, J. (2000). Exponential family models and statistical genetics. *Stat. Meth. Med. Res.* **9**, 57-72.

Paradis, E., Baillie, S. R., Sutherland, W. J. & Gregory, R. D. (1999). Dispersal and spatial scale affect synchrony in spatial population dynamics. *Ecol. Lett.* **2**, 114-120.

Peters, R. H. (1983). *The ecological implications of body size*. Cambridge: Cambridge University Press.

Purvis, A. & Garland, T., Jr. (1993). Polytomies in comparative analyses of continuous characters. *Syst. Biol.* **42**, 569-575.

Read, A. F. & Nee, S. (1995). Inference from binary comparative data. *J. theor. Biol.* **173**, 99-108.

Rohlf, F. J. (2001). Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution* **55**, 2143-2160.

Schmidt-Nielsen, K. (1984). *Scaling: why is animal size important?* Cambridge: Cambridge University Press.

Sibley, C. G. & Ahlquist, J. E. (1990). *Phylogeny and classification of birds: a study in molecular evolution*. New Haven: Yale University Press.

Sillén-Tullberg, B. (1993). The effect of biased inclusion of taxa on the correlation between discrete characters in phylogenetic trees. *Evolution* **47**, 1182-1191.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

APPENDIX

The generalized estimating equations in equation (3) can be solved through an iterative process which can be summarized as follows:

(i) compute an initial estimate of $\boldsymbol{\beta}$, for example, with a GLM;

(ii) compute an estimate of the variance-covariance matrix using equation (2);

(iii) update $\boldsymbol{\beta}$ with:

$$\boldsymbol{\beta}_{\text{step}+1} = \boldsymbol{\beta}_{\text{step}} - \left[\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right)^{\mathrm{T}} V^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right]^{-1} \left[\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right)^{\mathrm{T}} V^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right];$$

(iv) alternate between steps (ii) and (iii) until convergence.

Liang & Zeger (1986) introduced a robust (or empirical) estimator of W the variance-covariance matrix of $\boldsymbol{\beta}$. It is robust in the sense that it is consistent (i.e. it convergences to the true value of W when sample size increases) even if the correlation matrix R is misspecified. However, this robust estimator has very poor properties when the number of independent clusters is small (< 20), so it is not appropriate to the analysis of comparative data where there is a single cluster since all species are linked by their phylogenetic relationships. We prefer the use of a naive (or model-based) estimator of W given by:

$$\widehat{W}_{\text{naive}} = \left[\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right)^{\mathrm{T}} V^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right]^{-1}.$$

This estimator is consistent if R is correctly specified; however, Horton & Lipsitz (1999) report that it has better statistical properties than the robust estimator when the number of independent clusters is small even if R is wrong. Furthermore, we were able to verify that the naive estimator has good properties for continuous characters assuming a Gaussian response (see our present simulation study). However, it performs poorly for discrete characters assuming a binomial response since the number of independent clusters is too small (Mancl & DeRouen, 2001). We chose for binomial responses a quasi-likelihood estimator of W given by:

$$\widehat{W}_{\text{quasi}} = \left[-\frac{\partial^2 \ln Q}{\partial \boldsymbol{\beta}^2} \right]^{-1},$$

where Q is the quasi-likelihood function of the sample defined by (see Wedderburn, 1974):

$$\frac{\partial Q}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi \mu_i (1 - \mu_i)}.$$

There is in fact a close connection between GEE and quasi-likelihood (Liang & Zeger, 1986, p. 21).

Programs written in R (Ihaka & Gentleman, 1996), a freeware dialect of the S language (Becker *et al.*, 1988), are available from the authors.

Table 1. Results of analyses of simulated continuous characters along the phylogenies on Fig. 2. The parameter γ specifies the strength of the association between the two correlated characters. The figures in the table are the rates of rejection of the hypothesis of a null slope between the characters.

phylogeny	method	uncorrelated*	correlated characters (with $\gamma =$)†			
		characters	0.001	0.005	0.01	0.015
A	GEE	0.066	0.083	0.153	0.372	0.568
	contrasts	0.059	0.067	0.152	0.406	0.653
B	GEE	0.067	0.072	0.167	0.478	0.706
	contrasts	0.047	0.051	0.153	0.493	0.752
C	GEE	0.071	0.067	0.154	0.362	0.570
	contrasts	0.042	0.061	0.121	0.396	0.659

* In the case of uncorrelated evolution, the rejection rate is the type I error rate of the test.

† In the case of correlated evolution, the rejection rate is the power of the test ($= 1 -$ type II error rate).

Table 2. Results of analyses of simulated discrete characters along the phylogenies on Fig. 2.

phylogeny	method	$p = 0.6$		$p = 0.8$	
		uncorrelated	correlated	uncorrelated	correlated
A	GEE	0.040	0.057	0.049	0.370
	contrasts	0.087	0.095	0.092	0.513
B	GEE	0.040	0.028	0.041	0.371
	contrasts	0.078	0.068	0.073	0.536
C	GEE	0.073	0.084	0.088	0.369
	contrasts	0.128	0.133	0.129	0.538

Table 3. Analysis with generalized estimating equations of the effects of dispersal, habitat, and population trend on population synchrony of birds: tests of the significance of the effects.

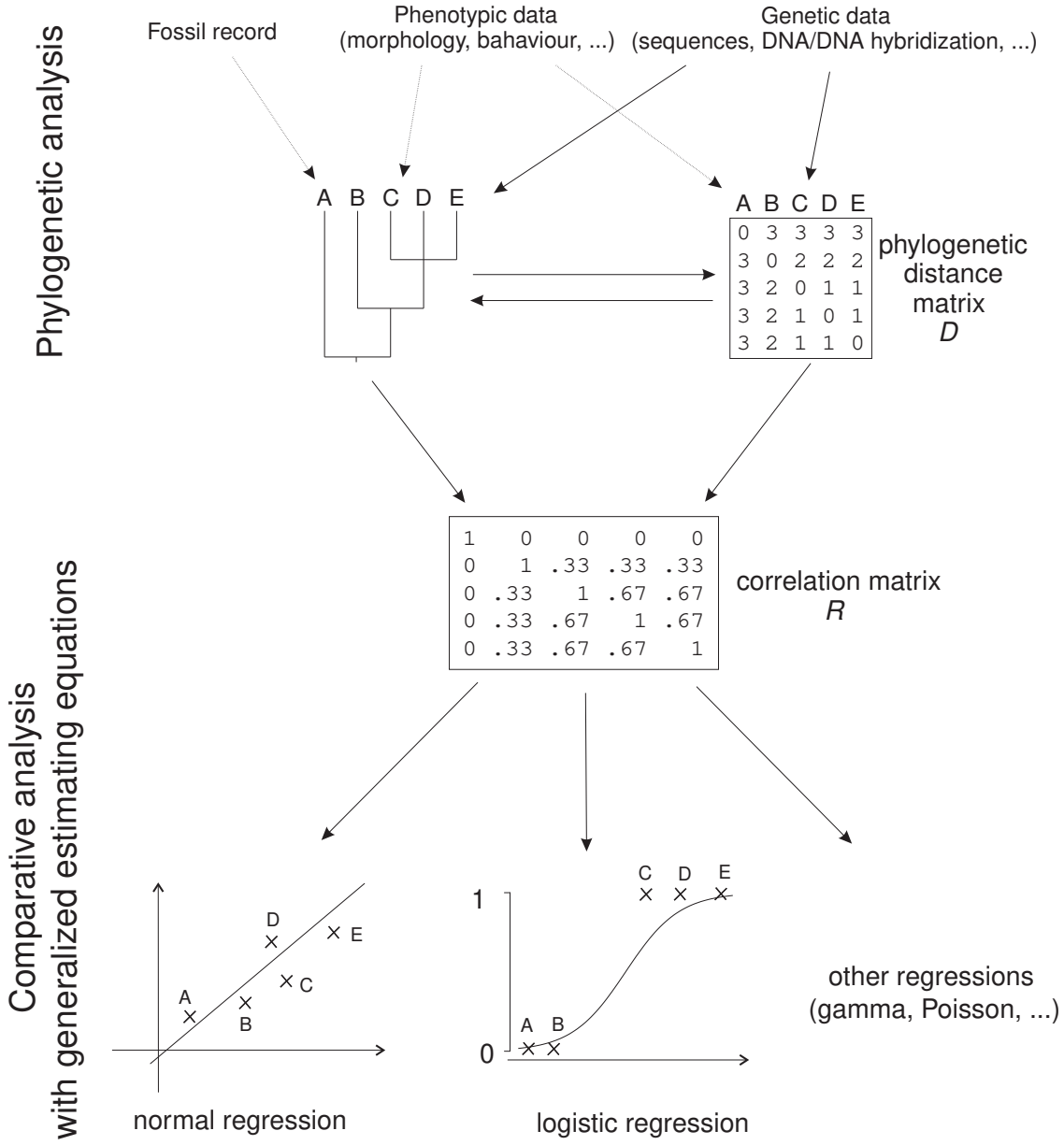
Dispersal	effect	F	df	P
Natal	full model	29.140	3,12	0.0001
	dispersal*habitat	27.473	2,12	0.0001
	long-term trend	5.189	1,12	0.042
Breeding	full model	29.018	3,12	0.0001
	dispersal*habitat	24.359	2,12	0.0001
	long-term trend	5.292	1,12	0.040

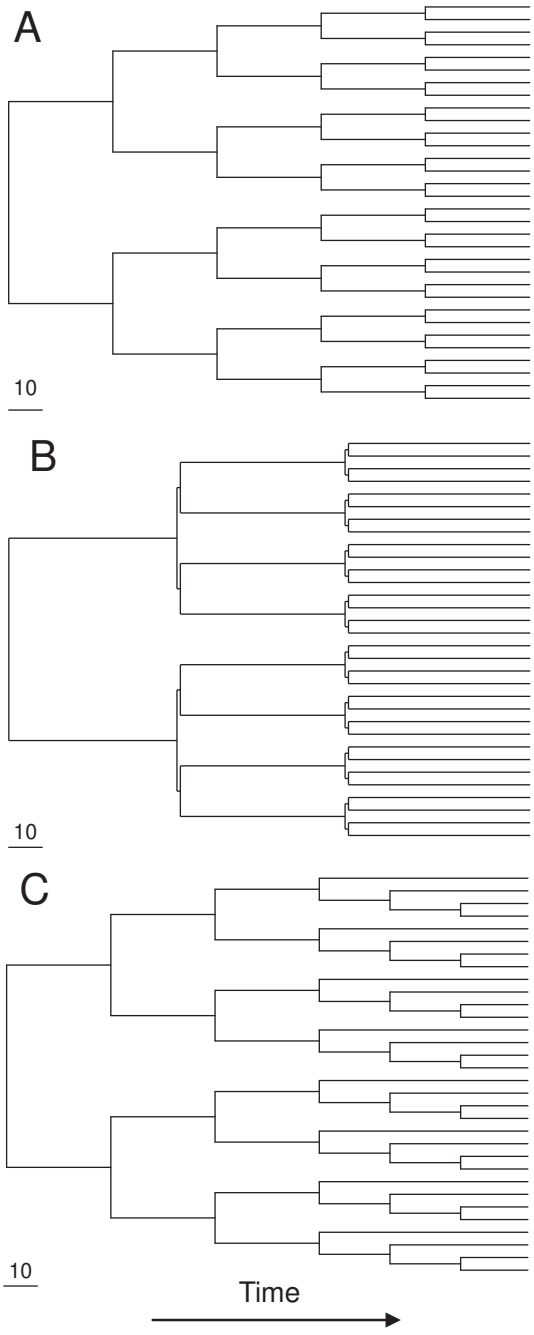
Table 4. Analysis with generalized estimating equations of the effects of dispersal, habitat, and population trend on population synchrony of birds: parameter estimates (SE: standard-error).

	natal dispersal		breeding dispersal	
	estimate	SE	estimate	SE
intercept	0.0926	0.0275	0.0937	0.0255
dispersal (dry habitats)	0.0073	0.0021	0.0099	0.0032
dispersal (wet habitats)	0.001	0.002	0.0017	0.003
long-term trend	−0.0003	0.0001	−0.0003	0.0001

Fig. 1. Schematic representation of the comparative analysis method using generalized estimating equations as proposed in the present paper.

Fig. 2. The three phylogenies used in the simulation study.


Phylogenetic analysis
Fossil record
Phenotypic data
(morphology, bahaviour, ...)
Genetic data
(sequences, DNA/DNA hybridization, ...)
A B C D E
A B C D E
0 3 3 3 3
3 0 2 2 2
3 2 0 1 1
3 2 1 0 1
3 2 1 1 0
phylogenetic
distance
matrix
D
1 0 0 0 0
0 1 .33 .33 .33
0 .33 1 .67 .67
0 .33 .67 1 .67
0 .33 .67 .67 1
correlation matrix
R
Comparative analysis
with generalized estimating equations
A
B
C
D
E
normal regression
1
0
A B
C D E
logistic regression
other regressions
(gamma, Poisson, ...)


A
10
B
10
C
10
Time