General Quantitative Genetic Methods for Comparative Biology

Chapter · July 2014

DOI: 10.1007/978-3-662-43550-2_11

Chapter 11
General Quantitative Genetic Methods for Comparative Biology

Pierre de Villemereuil and Shinichi Nakagawa

Abstract There is much in common between the aim and tools of the quantitative geneticist and the comparative biologist. One of the most interesting statistical tools of the quantitative genetics (QG) is the mixed model framework, especially the so-called animal model, which can be used for comparative analyses. In this chapter, we describe the phylogenetic generalised linear mixed model (PGLMM), which encompasses phylogenetic (linear) mixed model (PMM). The widely used phylogenetic generalised least square (PGLS) can be seen as a special case of PGLMM. Thus, we demonstrate how PGLMM can be a useful extension of PGLS, hence a useful tool for the comparative biologist. In particular, we show how the PGLMM can tackle issues such as (1) intraspecific variance inference, (2) phylogenetic meta-analysis, (3) non-Gaussian traits analysis, and (4) missing values and data augmentation. Further possible extensions of the PGLMM and applications to phylogenetic comparative (PC) analysis are discussed at the end of the chapter. We provide working examples, using the R package MCMCglmm, in the online practical material (OPM).

11.1 Introduction

Quantitative genetics (QG) and phylogenetic comparative (PC) methods have a lot in common, yet the connections between the two fields have only recently been stressed (Felsenstein 2005; Hadfield and Nakagawa 2010; Stone et al. 2011).

P. de Villemereuil (✉)
Laboratoire d'Écologie Alpine (LECA-UMR CNRS 5553), Université Joseph Fourier, BP 53, 38041 Grenoble, France
e-mail: bonamy@horus.ens.fr

S. Nakagawa
Department of Zoology, University of Otago, 340 Great King Street, Dunedin 9054, New Zealand
e-mail: shinichi.nakagawa@otago.ac.nz

L. Z. Garamszegi (ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, DOI: 10.1007/978-3-662-43550-2_11,

Indeed, both frameworks share many characteristics: (1) they aim at the evolutionary study of complex physiological, morphological, or ecological characters for which (2) they assume a Gaussian distribution (but see Sect. 11.3.1 for this assumption to be relaxed), and overall, (3) they aim at compartmentalising the phenotypic variability into an evolutive genetic component and one or several environmentally driven components. Quantitative geneticists have a long history of developing flexible and powerful statistical tools (see Hill and Kirkpatrick 2010, for a historical review), including the so-called 'animal model', which led to statistical developments such as the restricted maximum likelihood (REML) and the framework of (generalised) linear mixed models. Just as comparative phylogeny is using the relationship between species to investigate evolutionary events, the quantitative geneticists are interested into the relationship between individuals to infer the genetic component of polygenic traits. In particular, the 'animal model' is using a pedigree (a comprehensive record of the genealogy of the individuals) to decompose the phenotypic variance into its genetic and environmental components. To do so, the pedigree is transformed into a variance–covariance matrix of relatedness between individuals, which is included as a 'random effect' into the model. We will examine in this chapter how QG tools, namely (generalised) linear mixed models, can be adapted to the PC analysis framework and how they can nicely complement the widely used phylogenetic generalised least square (PGLS; for details of PGLS, see Chaps. 5 and 6). We explain that PGLS can, in fact, be seen as a special case of the phylogenetic (linear) mixed model (PMM) (Lynch 1991), which is, in turn, part of the overarching framework, phylogenetic generalised linear mixed model (PGLMM) (Hadfield and Nakagawa 2010; Ives and Helmus 2011). The evolutionary questions addressed in this chapter will thus be much alike those of the other chapters of Part II.

11.1.1 A Very Brief History of Phylogenetic Mixed Models

Lynch (1991) was the first to recognise the possibility to apply QG methods to comparative analysis, using mixed models to infer phylogeny-wide genetic variances against taxon-specific residual variance. The idea was to replace the variance–covariance matrix of relatedness between individuals by a phylogenetic variance–covariance (or correlation) matrix, which assumes Brownian motion model of trait evolution (i.e. assuming a constant variance in a trait through evolution, so that related species share closer trait values). By doing so, we could estimate ancestral states (or phylogenetic effects) instead of breeding values, and phylogenetic signal (the so-called phylogenetic heritability) instead of pedigree-based heritability Lynch (1991). Despite acknowledged interesting features (e.g. see Miles and Dunham 1993), Lynch's method PMM has only sparsely been highlighted in the PC literature (Housworth et al. 2004; Felsenstein 2008) and it seems to have rarely been used for practical comparative analysis. There are numerous reasons why this is the case. We can, however, come up with two

possible main reasons. First, Felsenstein's (1985) independent contrast (PIC) method had already set a standard for how to analyse inter-species comparative data before Lynch's (1991) QG-based method. Like other biological and human processes, it is likely that a founder-takes-all type of phenomenon has been at work (e.g. Waters et al. 2013). In other words, historical inertia (analogous to phylogenetic inertia!) may have played a role in this neglect on Lynch's important work. Second, unlike PIC, efficient algorithms and easy-to-use implementations have not been available for Lynch's method (at least until recently), even though Housworth et al. (2004) provided some improvement in algorithms, which has especially made estimation for multiresponse (multivariate) models more reliable. The under-usage of PMM feels little ironic because PIC is also a special case of PMM (Housworth et al. 2004).

After two and half decades since Lynch's work, Hadfield and Nakagawa (2010) revived the connections between QG and PC methods by developing a fast computational method for the phylogenetic variance–covariance matrix and its inverse. They have shown how PMM can be implemented in existing R packages (R Development Core Team 2011) and BUGS (Lunn et al. 2000). By developing an MCMC algorithm, they have also extended PMM to PGLMM, which can deal with non-Gaussian characters such as traits following binomial, multinomial, or Poisson distributions. Notably, they have proposed multinomial logit mixed models for a PC method. Such multinomial mixed models have not been used in QG although common in econometrics and political science. They showed that this multinomial PGLMM would be useful for the evolution of multiple discrete traits such as colour polymorphisms (i.e., for example, one taxon having three colour morphs, red, while, and black; see also Sect. 11.2.1). Recently, a number of comparative studies have tackled non-Gaussian traits using the framework of PGLMM (e.g. Ross et al. 2013a, b; Cornwallis et al. 2010; Maklakov et al. 2011).

We believe that it is worthwhile knowing the essence of Hadfield and Nakagawa's algorithm, although it is a little technical, as it represents the key connection between their method and QG animal models. There is a striking similarity between the phylogenetic variance–covariance matrix (hereby noted $\mathbf{\Sigma}$) and the relatedness matrix (hereby noted $\mathbf{A}$). As stated above, the former represents relatedness among species and is obtained from a phylogenetic tree, whereas the latter represents relatedness among individuals and is obtained from a pedigree. As the matrix $\mathbf{A}$ plays a critical role in estimating additive genetic variance and thus heritability of traits of interest, $\mathbf{\Sigma}$ allows us to estimate the phylogenetic variance and thus phylogenetic signal (see Sect. 11.2). For statistical computation, rather than $\mathbf{A}$ and $\mathbf{\Sigma}$, we require their inverse matrices, $\mathbf{A^{-1}}$ and $\mathbf{\Sigma}^{-1}$, whose computation can be extremely slow or even sometimes infeasible (this problem becomes increasingly worse as a pedigree or a phylogeny gets larger). So, the efficient algorithms of animal models (Henderson 1976; Meuwissen and Luo 1992) use the additive genetic variance $\mathbf{S}$, which is an expanded version of $\mathbf{A}$. The matrix $\mathbf{S}$ includes 'missing parents' so that all individuals including ones that do not have parents in an original pedigree will have a set of two parents. Importantly and rather counter-intuitively,

the inverse matrix, $\mathbf{S}^{-1}$, can be computed in much less time than $\mathbf{A}^{-1}$. This inclusion of missing parents is analogous to including the ancestral nodes because a pedigree and a phylogeny share the basic graph structure (with the phylogeny not having fathers). Furthermore, a branch length between parent and child node in a phylogeny is equivalent to inbreeding coefficient represented by a path between two individuals in a pedigree. Therefore, the phylogenetic version of $\mathbf{S}$, say $\boldsymbol{\Omega}$, can be constructed by including all ancestral nodes (not just tips, i.e. species), and the inverse of this (i.e. $\boldsymbol{\Omega}^{-1}$) can be used for computation. For example, with a large phylogeny (*ca.* 5,000 species), analysis with $\boldsymbol{\Sigma}^{-1}$ parametrisation (only using tips) could take over a month while the same analysis with $\boldsymbol{\Omega}^{-1}$ parametrisation (using tips and nodes) would only be a matter of an hour or so (for more technical details, see Hadfield and Nakagawa 2010).

11.1.2 Roadmap

In this chapter, we will show how QG methods can be useful for (1) multiple measurements data and intraspecific variance inference, (2) phylogenetic meta-analysis framework, (3) PC analysis on non-Gaussian characters, and (4) missing species design, using the framework of missing data theory. The chapter will end with a discussion about the interests and perspective of connections between QG and PC analysis frameworks. Although the sections of this chapter are quite independent from each other, readers who are unfamiliar with mixed models are strongly advised to read the following section. Also, it is recommended reading the following sections in the order of appearance. The reader will find working examples in the online practical material (hereafter OPM) at http://www.mpcm-evolution.org. The two most popular softwares for phylogeny-compatible mixed modelling are the frequentist software ASReml (Gilmour et al. 2006) and the Bayesian R package MCMCglmm (Hadfield 2010). Although the former is much faster, the OPM focus on the second package for several reasons. To begin with, MCMCglmm being Bayesian, it is more flexible than its frequentist equivalent and, in particular, it has better properties regarding non-Gaussian traits (de Villemereuil et al. 2013). Perhaps most importantly, the syntax of MCMCglmm is more oriented towards PC analysis and Hadfield and Nakagawa's (2010) algorithm has been directly implemented in it.

11.2 First Step: Mixed Model for Multiple Measurement Data

Random effects are commonly used within the mixed models framework to account for non-independent structure in the 'residuals'. In the context of comparative analysis, it can be useful to use such random effects to take phylogenetic

relationship between species into account. This section will constitute an introduction to mixed models and their applications to comparative analysis, by using the common case of multiple measurement data and intraspecific variance inference. For theoretical developments and review of methods for intraspecific variability, please refer to Chap. 7.

11.2.1 Description of the Simple Model

Let us assume we have phenotypic data **y** (e.g. body size) for several species and co-factors of interest (we will assume just one called **x**, e.g. the temperature of the environment). Now, consider we also have a phylogeny from which we derived a phylogenetic correlation matrix $\mathbf{\Sigma}$ (say using the classical Brownian motion assumption[1]). How can we define a mixed model to infer a relationship between **y** and **x** while taking the phylogenetic structure into account? The model would be as follows:

$$\mathbf{y} = \mu + \beta\mathbf{x} + \mathbf{a} + \mathbf{e} \tag{11.1}$$

where μ and β, respectively, are the intercept and the slope for the co-factor[2] **x**, **a** is the phylogenetic random effect, and **e** is the residual error. Now, the two last terms are assumed to be normally distributed with:

$$\begin{aligned} \mathbf{a} &\sim \mathcal{N}\left(0, \sigma_P^2 \mathbf{\Sigma}\right) \\ \mathbf{e} &\sim \mathcal{N}\left(0, \sigma_R^2 \mathbf{I}\right) \end{aligned} \tag{11.2}$$

where **I** stands for the relevant identity matrix. Our model, thus, assumes that phylogenetic effects are correlated according to the phylogenetic correlation matrix $\mathbf{\Sigma}$. Note also that our model is estimating two variances: V_P is the variance of the phylogenetic effect and V_R is the residual error (environment effects, intraspecific variance, measurement error, etc.).

It is important here to stress the resemblances and dissimilarities between the PGLMM above, and the classical model assumed in PGLS is denoted as:

$$\mathbf{y} \sim \mathcal{N}\left(\mu + \beta\mathbf{x}, \sigma_P^2 \mathbf{\Sigma}\right) \tag{11.3}$$

[1] But, any kind of evolutionary model yielding such a variance–covariance matrix can be used, such as Martins and Hansen's (1997) or ACDC processes (Blomberg et al. 2003). In practice, parameters of such models would be inferred before using the mixed model, but nothing, in theory, forbids the construction of a complex mixed model inferring these components along with performing the comparative regression.

[2] Of course, there can be an arbitrary number of such co-factors (either continuous or categorical variables).

Although the models are very much alike, a striking difference is the absence of the residual term e in the PGLS model, which only estimates σ^2_P, but not σ^2_R. In PC analysis, this constraint (that the residuals are distributed exactly according to the phylogeny) is usually relaxed using phylogenetic signal inference and introducing an extra parameter whose role is to measure such signal. By contrast, quantitative geneticists always assume that the pedigree (hence, the genetics) is only one source of the observed variability, the other one being the environment, usually captured by the residuals. Fortunately, comparative biologists do not have to give up on their usual tools to consider using mixed models. The model described in Eqs. 11.1 and 11.2 is equivalent to Pagel's λ model of phylogenetic signal inference (Freckleton et al. 2002; Housworth et al. 2004), given that the matrix $\mathbf{\Sigma}$ is a correlation matrix (i.e. diagonal elements are equal to 1, Hansen and Orzack 2005; Hadfield and Nakagawa 2010). Indeed, very much alike the heritability for QG analysis, we can define Lynch's phylogenetic heritability $\lambda = \frac{\sigma^2_P}{\sigma^2_P + \sigma^2_R}$ as a measure of the phylogenetic signal.[3] Actually, the above difference between PGLMM and PGLS is the only major one. Most other differences actually lie on which extensions of this model are used. For example, random effects and hierarchical modelling for non-Gaussian traits (see Sect. 11.3.1) are widely used in the field of QG, but scarcely in PC analyses. This chapter, among other things, aims at demonstrating how some of the quantitative geneticist 'tools' can prove to be useful to the comparative biologist.

11.2.2 Using Random Effects: The Case of Multiple Measurements

In many comparative cases, we have multiple measurements for each species. An extension to deal with such cases is straightforward, and we have:

$$\mathbf{y} = \mu + \beta\mathbf{x} + \mathbf{a} + \mathbf{s} + \mathbf{e} \tag{11.4}$$

$$\mathbf{s} \sim \mathcal{N}(0, \sigma^2_S \mathbf{I}) \tag{11.5}$$

where $\mathbf{s}$ is the 'multiple measurement effect' or species-specific effect after taking out the phylogenetic effect. This effect accounts for the variability that has been

[3] Note that, although λ could be forced to one by setting up $\sigma^2_R = 0$ in the model, this could cause numerical instability in frequentist software or strong auto-correlation in MCMC algorithms. The software MCMCglmm, for example, does not allow such a setting. Furthermore, there is some relevance in assuming that some of the biological variability is not captured by the phylogeny (such as environment or even measurement variability), hence assuming a residual variance. Also, notably, when $\sigma^2_R = 0$, PMM can be seen as equivalent to PGLS and thus PIC (Stone et al. 2011; Blomberg et al. 2012).

caused by the species' contingent characteristics (or species-specific effects). σ^2_S is the variance of this effect. The other symbols are as in Eqs. 11.1 and 11.2.

Together, σ^2_P and σ^2_S accounts for the between-species variability (the first being caused by the evolutionary history, the second by contingent events). By contrast, the residual term σ^2_R is a measure of the intraspecific variance of the trait.[4] Note that we are assuming the same intraspecific variance for all the species in the dataset, which might be considered as a very strong (although practical) assumption.

A careful inspection of Eq. 11.4 might reveal a troubling fact. As it stands, we have no clue of the type of relationship the slope β is measuring. As a comparative biologist, the reader would most likely be interested in the between-species slope. If the co-factor $\mathbf{x}$ only contains only one value per species (or mean specific values), then there is no problem, since for an individual j belonging to species i, the Eq. 11.4 can be rewritten as follows:

$$y_{i,j} - a_i - s_i = \mu + \beta x_i + e_{i,j} \tag{11.6}$$

Hence, we can consider the random effect a_i and s_i as within-species centring effects and the slope β as a between-species slope.

Things are slightly more complicated using individual measurements in $\mathbf{x}$, but it is still possible to obtain the between-species and within-species slopes using a technique called *within-group centring* (Davis et al. 1961; van de Pol and Wright 2009). The principle of this technique is to separate the predictor $\mathbf{x}$ into two components: one containing the group-level mean of $\mathbf{x}$ (here, the specific mean) and a second one containing the within-group variability. For an individual j belonging to species i, the new model would thus be:

$$y_{i,j} = \mu + \beta_B \bar{x}_i + \beta_W (x_{i,j} - \bar{x}_i) + a_i + s_i + e_{i,j} \tag{11.7}$$

where:

$$\bar{x}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} x_{i,j} \tag{11.8}$$

for J_i being the number of individuals in species i. Here, we are thus fitting two slopes: β_B is the slope of regression between species and β_W is the (common) slope of regression within each species. The model could be further complicated to include one slope per species (a so-called *random slope model*), but such a complex model would be out of the scope of this chapter. Note that, by construction, the predictors $\bar{x}_i$ and $(x_{i,j} - \bar{x}_i)$ are perfectly orthogonal. Therefore, β_B and β_W are truly independent. Finally, the calculation of λ would be changed to account for the extra random effect:

[4] This is not totally true, since σ^2_R also include noise such as measurement error, which is very difficult to distinguish from intraspecific variance without a careful design.

$$\lambda = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_S^2 + \sigma_R^2} \tag{11.9}$$

We are thus able to estimate intraspecific variance and between-species slope for multiple measurement data with the help of a new random effect. This is only a particular demonstration of the utility of multiple random effects model. One could also use them to account for problems of unbalanced sampling in one's dataset: spatial correlation, biogeographic regions, etc. (see Ives and Zhu 2006). In theory, most of the dependency structures in the error of the model could be accounted for by a random effect.

11.2.3 Phylogenetic Meta-Analysis Using Random Effects

Meta-analysis is a powerful statistical tool to combine weighted results of multiple studies on the same or similar topics. As such, although the technique originated from medical and social sciences, meta-analysis has been used extensively in the field of ecology and evolution (Nakagawa and Poulin 2012; Koricheva et al. 2013). In ecological or evolutionary meta-analysis, it is common that data include multiple species, and therefore, the data look similar to those of comparative analysis. The main difference is that what are 'traits' in PC analysis (e.g. brain size) are 'effect sizes' in meta-analysis (e.g. a relationship between brain size and reproductive success within a species). Such effect sizes are commonly standardised statistical metrics, which are dimensionless (Cohen 1988; Nakagawa and Cuthill 2007), so that they can be compared across studies or species. Four commonly used effect size metrics[5] are: (1) Fisher's z-transformation of correlation coefficient (*Zr*), (2) Hedges' *d* and its variants, (3) response ratio on the natural logarithm (ln*R*), and (4) odds ratio on the natural logarithm (ln*OR*) (Nakagawa and Santos 2012; Koricheva et al. 2013). A recent study suggests the importance of incorporating phylogeny in meta-analysis because meta-analytic models with and without phylogeny could result in different conclusions (Chamberlain et al. 2012). Here, we will describe phylogenetic meta-analytic models. A working example of such analysis can be found in the OPM.

Several versions of phylogenetic meta-analysis have been proposed (Adams 2008; Lajeunesse 2009; Hadfield and Nakagawa 2010). Although they are slightly different in their details, they all aim for incorporating phylogenetic non-independence. Here, we describe the one based on PMM, described in Hadfield and Nakagawa (2010). In a phylogenetic meta-analytic, we have a vector of effect sizes $\mathbf{z}$ and each effect size has its sampling error variance (all stored in a vector $\mathbf{v_m}$),

[5] These standardised metrics are unbounded and follow approximately normal distributions. However, note that the correlation coefficient *r* is bounded at −1 and 1 and does not follow a normal distribution.

which may sometimes be referred to as measurement error variance. The model is denoted as:

$$\mathbf{z} = \mu + \mathbf{a} + \mathbf{m} + \mathbf{e} \tag{11.10}$$

$$\mathbf{m} \sim \mathcal{N}(0, \mathbf{v_m I}) \tag{11.11}$$

where μ is the meta-analytic (grand) mean, $\mathbf{m}$ are the sampling (measurement) error effects,[6] and all the other symbols are as in Eqs. 11.1 and 11.2. Sampling error variance is assumed to be known, and for all common effect size statistics, equations are available to obtain their sampling error variances. For example, the sample variance for Zr is $\frac{1}{n-3}$, where n is sample size used to estimate a correlation coefficient. Also note that meta-analysis is typically an intercept model (i.e. μ is the only fixed factor estimated). This is because the main purpose of a meta-analysis is to identify a general trend. But, as you may realise, it is easy to add a predictor (co-factor $\mathbf{x}$ in Eq. 11.1), by building up on Eq. 11.10. This model is expressed as:

$$\mathbf{z} = \mu + \beta\mathbf{x} + \mathbf{a} + \mathbf{m} + \mathbf{e}. \tag{11.12}$$

This model is known as (phylogenetic) meta-regression, and we can add as many fixed factors and random factors as required. Such mixed model-based phylogenetic meta-analysis or meta-regression has recently been used in a number of studies (e.g. Horváthová et al. 2012; Santos and Nakagawa 2012; Prokop et al. 2012; Garamszegi et al. 2012).

Incidentally, remember the example in Sect. 11.2. There, we had multiple measurements per species. Rather than all these raw multiple measurements, let us suppose that we only have species trait means and standard errors (or alternatively, standard deviations and sample sizes). In such a case, we have a model where the square of standard errors for each species trait value can be considered as sampling error variances $\mathbf{v_m}$ so that:

$$\mathbf{y} = \mu + \beta\mathbf{x} + \mathbf{a} + \mathbf{m} + \mathbf{e} \tag{11.13}$$

where $\mathbf{y}$ is a vector of the trait mean for each species. Now, we can see that a phylogenetic meta-regression (Eq. 11.12) and a PC analysis incorporating within-species (sampling error) variance (Eq. 11.13) are mathematically equivalent (see Chap. 7); such equivalence has been pointed out previously (Nakagawa and Santos 2012; Jennions et al. 2012).

[6] In a typical non-phylogenetic meta-analysis, a unit of analysis is 'study' where one effect size is taken from one study. Here, we assume that one effect size from each species comes from one study or $n_{\text{effect}} = n_{\text{species}} = n_{\text{study}}$.

11.3 Extensions of the Phylogenetic Linear Mixed Model

11.3.1 Non-Gaussian Characters: Generalised Linear Mixed Model

One of the main advantages of the mixed model framework is that it has been generalised to non-Gaussian response distribution (Gilmour et al. 1985; Breslow and Clayton 1993). As a result, it is now relatively easy to investigate non-Gaussian comparative data using a generalised phylogenetic mixed model (although one needs to be aware of the classical pitfalls, see Bolker et al. 2009). Because of the flexibility of the MCMC algorithm, the MCMCglmm package is one of the most comprehensive in terms of the available distributions (even some complex ones like zero-inflated Poisson). Note that non-Gaussian traits can also be analysed using ASReml software with common distributions such as Poisson and binomial.[7]

Mixed models are generalised by adding a hierarchical layer in the model described in Eq. 11.1. Indeed, we begin by assuming a hypothetical latent trait $\mathbf{l}$, which satisfies Eq. 11.1:

$$\mathbf{l} = \mu + \beta\mathbf{x} + \mathbf{a} + \mathbf{e} \tag{11.14}$$

Note that all assumptions detailed in Sect. 11.2 hold for $\mathbf{l}$ here, since we are using the same model.[8] Then, we use a 'link function' g to draw the relationship between this latent trait $\mathbf{l}$ and our actual non-Gaussian data $\mathbf{y}$. This function will transform $\mathbf{l}$ into a quantity $g^{-1}(\mathbf{l})$, which will be the expectation of the distribution of $\mathbf{y}$.

For example, for count data, we will assume a Poisson distribution (noted $\mathscr{P}$), which only accepts positive mean. The canonical link function is the logarithm, so that:

$$\mathbf{y} \sim \mathscr{P}(\exp(\mathbf{l})) \tag{11.15}$$

You will find a working example in the OPM using count data and Poisson distribution.

For dichotomous (binary) data, we will assume a binomial distribution (noted $\mathscr{B}$) with a vector of probability of "success", $\mathbf{p}$:

$$\mathbf{y} \sim \mathscr{B}(\mathbf{p}) \tag{11.16}$$

[7] However, the penalised quasi-likelihood used in ASReml has been shown to largely underestimate the variance components for binary traits (Gilmour et al. 2006; de Villemereuil et al. 2013).

[8] Note, however, that $\mathbf{e}$ in Eq. 11.14 should be considered as the effect due to additive dispersion rather than the residuals (for additive dispersion, see Nakagawa and Schielzeth 2010).

One of the two link functions is usually assumed: the logit or the probit functions. The logit link function is the canonical link function, calculated as:

$$\mathbf{l} = \log\left(\frac{\mathbf{p}}{1-\mathbf{p}}\right) \tag{11.17}$$

The probit link function is actually the inverse function of the cumulative distribution function of a standard normal distribution, noted Φ:

$$\mathbf{l} = \Phi^{-1}(\mathbf{p}) \tag{11.18}$$

The probit has less mathematical support than the logit, but it has a better biological interpretation, due to its strong link with the long-standing *threshold model* (Wright 1934; Dempster and Lerner 1950), widely used in QG for dichotomous traits (see also Chap. 9 for the use of phylogenetic logistic regression for binary traits).

Because of this hierarchical modelling, the generalisation of the phylogenetic mixed model to non-Gaussian traits allows us to keep assumptions about the main Gaussian evolution processes, for example the Brownian motion model. Regarding dichotomous traits, this model is indeed philosophically different from the Markovian processes widely used in the field of PC analysis (see Chaps. 10 and 16). Whereas the latter model discretise the genetics of the dichotomous trait in assuming probabilities of 'jump' from one state to the other, the former assume a polygenic basis of the trait and thus a 'smoother' evolution (and possibly intraspecific variability). More generally, a Gaussian latent trait is justified if you consider highly polygenic characters (even for the discrete ones!), because then the result of all genetic effects should be normally distributed, according to Fisher's (1918) infinitesimal model. Another interest lies, of course, in the fact that most kinds of distributions can be used in Eq. 11.15. Indeed, most ecologically interesting distributions are available in the MCMCglmm package (binomial, ordinal, multinomial, Poisson, zero-inflated Poisson, etc.).

11.3.2 Missing Species and Data Augmentation

When we compile comparative data of a certain taxon, it is difficult to get all trait values for every species in that taxon. In other words, missing data are commonplace in comparative data. What most of researchers have been doing is to ignore species in which trait information is missing and to conduct analysis without such species—this is called 'complete-case analysis'. Unfortunately, complete-case analysis in comparative data will often result in biased estimates (Nakagawa and Freckleton 2008; Garamszegi and Møller 2011). To understand why this is so, we will benefit from learning some basics on missing data mechanisms.

Missing data mechanisms in the statistical literature are merely a classification of how missing data are related to observed data so that missing data mechanisms do not imply casual explanations for missing data. A series of work by Rubin and Little has set foundations for missing data theory (Rubin 1976, 1987; Little and Rubin 2002). In missing data theory, three types of missing data are recognised: (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) missing not at random (MNAR). To understand these mechanisms, we need to introduce three concepts, by using notations by Enders (2010). First, we need to recognise that the data matrix $\mathbf{Y}$ can be decomposed into an observed part $\mathbf{Y}_o$ and a missing part $\mathbf{Y}_m$. Second, the matrix $\mathbf{R}$ that has the same dimension as $\mathbf{Y}$ has either 0 or 1 in its elements, with 0 signifying 'observed' and 1 'missing'. The matrix $\mathbf{R}$ is referred to as missingness. Third, we call $\boldsymbol{\theta}$ the vector of parameters that described the relationships between the data $\mathbf{Y}$ and its missingness $\mathbf{R}$. Now, the probability distribution function for MCAR can be written as:

$$p(\mathbf{R}|\boldsymbol{\theta}) \tag{11.19}$$

It reads as follows: the probability of whether an element in $\mathbf{R}$ takes 0 or 1 depends neither on $\mathbf{Y}_o$ nor on $\mathbf{Y}_m$. The lack of links between $\mathbf{R}$ and the data $\mathbf{Y}$ is solely described by $\boldsymbol{\theta}$ (cf. Enders 2010). That is, missing values in a variable of interest are distributed completely at random in relation to any other variables. This function also implies that when missing data is MCAR, the complete-case analysis will provide unbiased results although statistical power may be reduced. However, MCAR is a very strong and actually unrealistic assumption because biological processes usually cause missing values. For example, we are less likely to have life-history data on rare species than on more abundant ones. Therefore, a more realistic assumption for missing data is MAR. The probability distribution function for MAR is:

$$p(\mathbf{R}|\mathbf{Y}_o, \boldsymbol{\theta}) \tag{11.20}$$

Again, it reads as follow: the probability of having 0 or 1 in $\mathbf{R}$ depends only on $\mathbf{Y}_o$, and this relationship between $\mathbf{R}$ and $\mathbf{Y}$ is described by $\boldsymbol{\theta}$. That is, missing values in a variable of interest are due to another variable that we have complete data on. For example, missing values in life-history data of rarer species are MAR, if we have complete abundance data, which governs the probability of missingness. It is notable that MAR does not really mean 'MAR' in its usual sense (although this might be very confusing).

Similarly, the probability distribution function for MNAR is:

$$p(\mathbf{R}|\mathbf{Y}_o, \mathbf{Y}_m, \boldsymbol{\theta}) \tag{11.21}$$

It reads as follow: the probability of taking 0 or 1 in $\mathbf{R}$ depends both on $\mathbf{Y}_o$ and $\mathbf{Y}_m$, and this relationship between $\mathbf{R}$ and $\mathbf{Y}$ is described by $\boldsymbol{\theta}$. That is, missing values in a variable of interest are due to another variable that we do not have any

information on or missing values are due to missing values themselves. For example, missing data in life-history data of rarer species are MNAR, if we do not have abundance data, which governs missingness, or if a life-history trait makes particular species rare so that it is difficult to obtain such life-history data of rare species. Importantly, if we conduct complete-case analysis on MAR or MNAR missing data, parameter estimates will often be biased. Therefore, all comparative analyses not accounting for missing species could provide biased parameters and potentially lead to incorrect conclusions, because it is unlikely that such missing species comply with the MCAR assumption. A working example using MCMCglmm that assumes MAR can be found in the OPM.

An important problem here is that we never really know how MNAR missing values were created because the patterns of missingness depend on missing values themselves. Thus, treating MNAR missingness is usually very difficult or often infeasible. Therefore, the most practical assumption is MAR, and many methods to deal with missing data under MAR have been developed (Enders 2010; van Buuren 2012). One notable method is data augmentation using Bayesian MCMC. Although methodological details are beyond the scope of this chapter (for an accessible account, see Enders 2010), data augmentation will provide unbiased parameter estimates when missing data in a dataset fulfils MAR. MCMCglmm uses a data augmentation method when missing data are in the response variable, but not predictors. Datasets with missing data in multiple variables need to resort to either using other methods such as multiple imputation (van Buuren 2012) or using multiresponse models, which ASReml and MCMCglmm are capable of running (for details, see Gilmour et al. 1985; Hadfield 2010). Although there are still a handful of examples of comparative studies utilising the missing data theory (e.g. Fisher et al. 2003; Gonzalez-Suarez et al. 2012; Cleasby and Nakagawa 2012), we expect that the use of missing data missing data augmentation or other methods for dealing with missing data such as multiple imputation will be commonplace in PC analysis in the near future.

11.4 Discussion

Throughout this chapter, we have seen some of the most interesting properties of PGLMM, arising from using QG mixed models for PC analysis. Because mixed models are nowadays a standard in ecology and evolution, and the (sometimes quite philosophical) difference between fixed and random effects is relatively well understood by the scientific community (reviewed in Gelman and Hill 2006), we expect that a shift towards phylogenetic mixed model will not be a huge step for any practitioner interested in evolution. Such a shift would not represent a significant change in the underlying model, since there is a strong proximity between phylogenetic mixed models and phylogenetic generalised least squares. Yet, the

GLMM framework might allow one to tackle some of the important issues listed in the sections above. Incidentally, note that all the examples in this chapter use one response variable, but all of the models above can be easily extended to be multiresponse models, for which MCMCglmm provides an implementation (Hadfield and Nakagawa 2010; Nakagawa and Santos 2012), as well as ASReml (Gilmour et al. 2006). Additionally, multiresponse models for non-Gaussian traits can account for a different distribution for each trait (Hadfield 2010).

Regarding the implementation of the models, we focused on the MCMCglmm package in this chapter. Yet, although the Bayesian MCMC algorithm possesses numerous advantageous sides in the case of GLMM implementation, it is important to stress that one does not need to be Bayesian to run such models.[9] However, we are still lacking a nice and flexible frequentist package that can readily be used in the context of PGLMM (but see Ives and Helmus 2011), mainly because accurate likelihood-based estimations for generalised linear mixed models have proven difficult and still are an active area of research in statistics.

Further extensions of the phylogenetic mixed model are to be expected in the future. For example, Stone et al. (2011) suggested, among other methods, the use of MCMCglmm to control for population structure within species. On the same idea, Buckley et al. (2010) used MCMCglmm to fit 'species within the phylogeny', 'populations within species', and 'years within population' effects. Such nested design is easy and quite intuitive to implement using random effects in a mixed model context (Schielzeth and Nakagawa 2013). Also, one can incorporate two different phylogenies as random effects to investigate traits whose evolution is affected by interactions between species, such as host–parasite, plant–pollinator, and predator–prey interactions (Rafferty and Ives 2013; Hadfield et al. 2014). Another way to extend the mixed model framework would be to allow for uncertainty in the relatedness matrix. In the context of PGLS, de Villemereuil et al. (2012) showed that failing to take this uncertainty into account could lead to anti-conservative standard error of estimates. Although technically difficult to implement efficiently in a package such as MCMCglmm (J. Hadfield, personal communication, but see Ross et al. 2013a, about an implementation of such uncertainty using MCMCglmm), sampling in a distribution of phylogenies instead of using the consensus, one might lead to better estimations (Huelsenbeck and Rannala 2003; de Villemereuil et al. 2012). Because pedigrees are also not known without error (either being social or genetically determined, see Charmantier and Réale 2005; Sillanpää 2011), such 'phylogenetic/animal models with uncertainty' would benefit both quantitative genetic and comparative analysis statistical fields.

Acknowledgments We are grateful for S. Lavergne, M. Lagisz, L. Z. Garamszegi and two anonymous reviewers for their comments on our earlier versions of this chapter; their comments have significantly improved this chapter. S.N. is supported by the Rutherford Discovery Fellowship (New Zealand).

[9] The data augmentation of Sect. 11.3.2, though, is very much linked to the MCMC algorithm.

References

Adams D (2008) Phylogenetic meta-analysis. Evolution. Int J Organ Evol 62(3):567–572. doi:10.1111/j.1558-5646.2007.00314.x

Blomberg SP, Garland T, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57(4):717–745

Blomberg SP, Lefevre JG, Wells JA, Waterhouse M (2012) Independent contrasts and PGLS regression estimators are equivalent. Syst Biol 61(3):382–391. doi:10.1093/Sysbio/Syr118

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24(3):127–135. doi:10.1016/j.tree.2008.10.008

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88(421):9–25. doi:10.2307/2290687

Buckley YM, Ramula S, Blomberg SP, Burns JH, Crone EE, Ehrlén J, Knight TM, Pichancourt JB, Quested H, Wardle GM (2010) Causes and consequences of variation in plant population growth rate: a synthesis of matrix population models in a phylogenetic context. Ecol Lett 13(9):1182–1197. doi:10.1111/j.1461-0248.2010.01506.x

Chamberlain S, Hovick S, Dibble C, Rasmussen N, Van Allen B, Maitner B, Ahern J, Lukas B, Roy C, Maria M, Carrillo J, Siemann E, Lajeunesse M, Whitney K (2012) Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. Ecol Lett 15(6):627–636. doi:10.1111/j.1461-0248.2012.01776.x

Charmantier A, Réale D (2005) How do misassigned paternities affect the estimation of heritability in the wild? Mol Ecol 14(9):2839–2850. doi:10.1111/j.1365-294X.2005.02619.x

Cleasby IR, Nakagawa S (2012) The influence of male age on within-pair and extra-pair paternity in passerines. Ibis 154(2):318–324. doi:10.1111/J.1474-919x.2011.01209.X

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum, Hillsdale, New Jersey

Cornwallis CK, West SA, Davis KE, Griffin AS (2010) Promiscuity and the evolutionary transition to complex societies. Nature 466(7309):969–72. doi:10.1038/nature09335

de Villemereuil P, Wells JA, Edwards RD, Blomberg SP (2012) Bayesian models for comparative analysis integrating phylogenetic uncertainty. BMC Evol Biol 12(1):102. doi:10.1186/1471-2148-12-102

de Villemereuil P, Gimenez O, Doligez B (2013) Comparing parent–offspring regression with frequentist and bayesian animal models to estimate heritability in wild populations: a simulation study for gaussian and binary traits. Meth Ecol Evol 4(3):260–275. doi:10.1111/2041-210X.12011

Davis J, Spaeth J, Huson C (1961) A technique for analyzing the effects of group composition. Am Sociol Rev 26(2):215–225. doi:10.2307/2089857

Dempster ER, Lerner IM (1950) Heritability of threshold characters. Genetics 35(2):212–236

Enders CK (2010) Applied missing data analysis. Methodology in the social sciences. Guilford Press, New York, 2010008465 GBB060973 Craig K. Enders. ill.; 26 cm. Includes bibliographical references (p 347–358) and indexes. Methodology in the social sciences

Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 1–15

Felsenstein J (2005) Using the quantitative genetic threshold model for inferences between and within species. Philos Trans: Biol Sci 360(1459):1427–1434

Felsenstein J (2008) Comparative methods with sampling error and within-species variation: contrasts revisited and revised. Am Nat 171(6):713–725. doi 10.1086/587525

Fisher D, Blomberg S, Owens I (2003) Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials. Proc Biol Sci/Roy Soc 270(1526):1801–1808. doi:10.1098/rspb.2003.2447

Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans Roy Soc Edinb 52:399–433

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: A test and review of evidence. Am Nat 160(6):712–726. doi:10.1086/343873

Garamszegi LZ, Møller AP (2011) Nonrandom variation in within-species sample size and missing data in phylogenetic comparative studies. Syst Biol 60(6):876–880

Garamszegi LZ, Marko G, Herczeg G (2012) A meta-analysis of correlated behaviours with implications for behavioural syndromes: mean effect size, publication bias, phylogenetic effects and the role of mediator variables. Evol Ecol 26(5):1213–1235. doi:10.1007/S10682-012-9589-8

Gelman A, Hill J (2006) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge

Gilmour AR, Anderson RD, Rae AL (1985) The analysis of binomial data by a generalized linear mixed model. Biometrika 72(3):593–599. doi:10.1093/biomet/72.3.593

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2006) ASReml user guide release 2.0. http://www.vsni.co.uk/software/asreml/

Gonzalez-Suarez M, Lucas PM, Revilla E (2012) Biases in comparative analyses of extinction risk: mind the gap. J Anim Ecol 81(6):1211–1222

Hadfield JD (2010) MCMC methods for multi-response generalised linear mixed models: The MCMCglmm R package. J Stat Softw 33(2):1–22

Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol 23(3):494–508. doi:10.1111/j.1420-9101.2009.01915.x

Hadfield JD, Kranov B, Poulin R, Nakagawa S (2014) A tale of two phylogenies: comparative analyses of ecological interactions. Am Nat 183(2):174–187

Hansen TF, Orzack SH (2005) Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: the need for controlled comparisons. Evolution 59(10):2063–2072

Henderson C (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 1:69–83

Hill WG, Kirkpatrick M(2010) What animal breeding has taught us about evolution. Ann Rev Ecol, Evol Syst 41:1–19. doi:10.1146/annurev-ecolsys-102209-144728

Horváthová T, Nakagawa S, Uller T (2012) Strategic female reproductive investment in response to male attractiveness in birds. Proc Roy Soc B-Biol Sci 279(1726):163–170

Housworth E, Martins E, Lynch M (2004) The phylogenetic mixed model. Am Nat 163(1):84–96. doi:10.1086/380570

Huelsenbeck JP, Rannala B (2003) Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution 57(6):1237–1247

Ives AR, Helmus MR (2011) Generalized linear mixed models for phylogenetic analyses of community structure. Ecol Monogr 81(3):511–525

Ives AR, Zhu J (2006) Statistics for correlated data: phylogenies, space, and time. Ecol Appl 16(1):20–32

Jennions MD, Kahn AT, Kelly CD, Kokko H (2012) Meta-analysis and sexual selection: past studies and future possibilities. Evol Ecol 26(5):1119–1151. doi: 10.1007/S10682-012-9567-1

Koricheva J, Gurevitch J, Mengersen K (2013) The handbook of meta-analysis in ecology and evolution. Princeton University Press, Princeton

Lajeunesse M (2009) Meta-analysis and the comparative phylogenetic method. The Am Nat 174(3):369–381. doi:10.1086/603628

Little RJA, Rubin DB (2002) Statistical analysis with missing data, Wiley series in probability and statistics, 2nd edn. Wiley, Hoboken, N.J., p 349–364

Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) Winbugs—a bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10(4):325–337

Lynch M (1991) Methods for the analysis of comparative data in evolutionary biology. Evolution 45(5):1065–1080. doi:10.2307/2409716

Maklakov AA, Immler S, Gonzalez-Voyer A, Ronn J, Kolm N (2011) Brains and the city: big-brained passerine birds succeed in urban environments. Biol Lett 7(5):730–732. doi:10.1098/Rsbl.2011.0341

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149(4):646–667

Meuwissen T, Luo Z (1992) Computing inbreeding coefficients in large populations. Genet Sel Evol 24:305–313. doi:10.1186/1297-9686-24-4-305

Miles DB, Dunham AE (1993) Historical perspectives in ecology and evolutionary biology: the use of phylogenetic comparative analyses. Ann Rev Ecol Syst 587–619

Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev 82(4):591–605

Nakagawa S, Freckleton RP (2008) Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol 23(11):592–596

Nakagawa S, Poulin R (2012) Meta-analytic insights into evolutionary ecology: an introduction and synthesis. Evol Ecol 26(5):1085–1099

Nakagawa S, Santos ESA (2012) Methodological issues and advances in biological meta-analysis. Evol Ecol 26(5):1253–1274

Nakagawa S, Schielzeth H (2010) Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biol Rev Camb Philos Soc 85(4):935–956. doi:10.1111/j.1469-185X.2010.00141.x

Prokop ZM, Michalczyk L, Drobniak SM, Herdegen M, Radwan J (2012) Metaanalysis suggests choosy females get sexy sons more than "good genes". Evolution 66(9):2665–2673

R Development Core Team (2011) {R}: a language and environment for statistical computing. http://www.R-project.org/

Rafferty NE, Ives AR (2013) Phylogenetic trait-based analyses of ecological networks. Ecology in press

Ross L, Gardner A, Hardy N, West SA (2013a) Ecology, not the genetics of sex determination, determines who helps in eusocial populations. Curr Biol 23(23):2383–2387. doi:10.1016/j.cub.2013.10.013

Ross L, Hardy NB, Okusu A, Normark BB (2013b) Large population size predicts the distribution of asexuality in scale insects. Evolution 67(1):196–206. doi:10.1111/J.1558-5646.2012.01784.X

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–590

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York, NY

Santos ESA, Nakagawa S (2012) The costs of parental care: a meta-analysis of the trade-off between parental effort and survival in birds. J Evol Biol 25(9):1911–1917. doi:10.1111/J.1420-9101.2012.02569.X

Schielzeth H, Nakagawa S (2013) Nested by design: model fitting and interpretation in a mixed model era. Meth Ecol Evol 4(1):14–24. doi:10.1111/j.2041-210x.2012.00251.x

Sillanpää MJ (2011) On statistical methods for estimating heritability in wild populations. Mol Ecol 20(7):1324–1332. doi:10.1111/j.1365-294X.2011.05021.x

Stone GN, Nee S, Felsenstein J (2011) Controlling for non-independence in comparative analysis of patterns across populations within species. Philos Trans Roy Soc B: Biol Sci 366(1569):1410 –1424. doi:10.1098/rstb.2010.0311

van Buuren S (2012) Flexible imputation of missing data. Chapman and hall/CRC interdisciplinary statistics series. CRC Press, Boca Raton, FL

van de Pol M, Wright J (2009) A simple method for distinguishing within- versus between-subject effects using mixed models. Anim Behav 77(3):753–758. doi:10.1016/j.anbehav.2008.11.006

Waters J, Fraser C, Hewitt G (2013) Founder takes all: density-dependent processes structure biodiversity. Trends Ecol Evol 28(2):78–85. doi:10.1016/j.tree.2012.08.024

Wright S (1934) An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics 19(6):506–536