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,

might be all the cholesky stuff etc.

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 $\mathbf{\Omega}$, can be constructed by including all ancestral nodes (not just tips, i.e. species), and the inverse of this (i.e. $\mathbf{\Omega}^{-1}$) can be used for computation. For example, with a large phylogeny (*ca.* 5,000 species), analysis with $\mathbf{\Sigma}^{-1}$ parametrisation (only using tips) could take over a month while the same analysis with $\mathbf{\Omega}^{-1}$ 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 $\mathbf{y}$ (e.g. body size) for several species and co-factors of interest (we will assume just one called $\mathbf{x}$, e.g. the temperature of the environment). Now, consider we also have a phylogeny from which we derived a phylogenetic correlation matrix $\boldsymbol{\Sigma}$ (say using the classical Brownian motion assumption[1]). How can we define a mixed model to infer a relationship between $\mathbf{y}$ and $\mathbf{x}$ 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] $\mathbf{x}$, $\mathbf{a}$ is the phylogenetic random effect, and $\mathbf{e}$ is the residual error. Now, the two last terms are assumed to be normally distributed with:

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

where $\mathbf{I}$ stands for the relevant identity matrix. Our model, thus, assumes that phylogenetic effects are correlated according to the phylogenetic correlation matrix $\boldsymbol{\Sigma}$. 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:

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

[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_P^2, but not σ_R^2. 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 $\boldsymbol{\Sigma}$ 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 $\lambda = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_R^2}$ 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:

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

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

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 $\sigma_R^2 = 0$ 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 $\sigma_R^2 = 0$, 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.

$$\lambda = \frac{\sigma^2_P}{\sigma^2_P + \sigma^2_S + \sigma^2_R} \tag{11.9}$$

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.