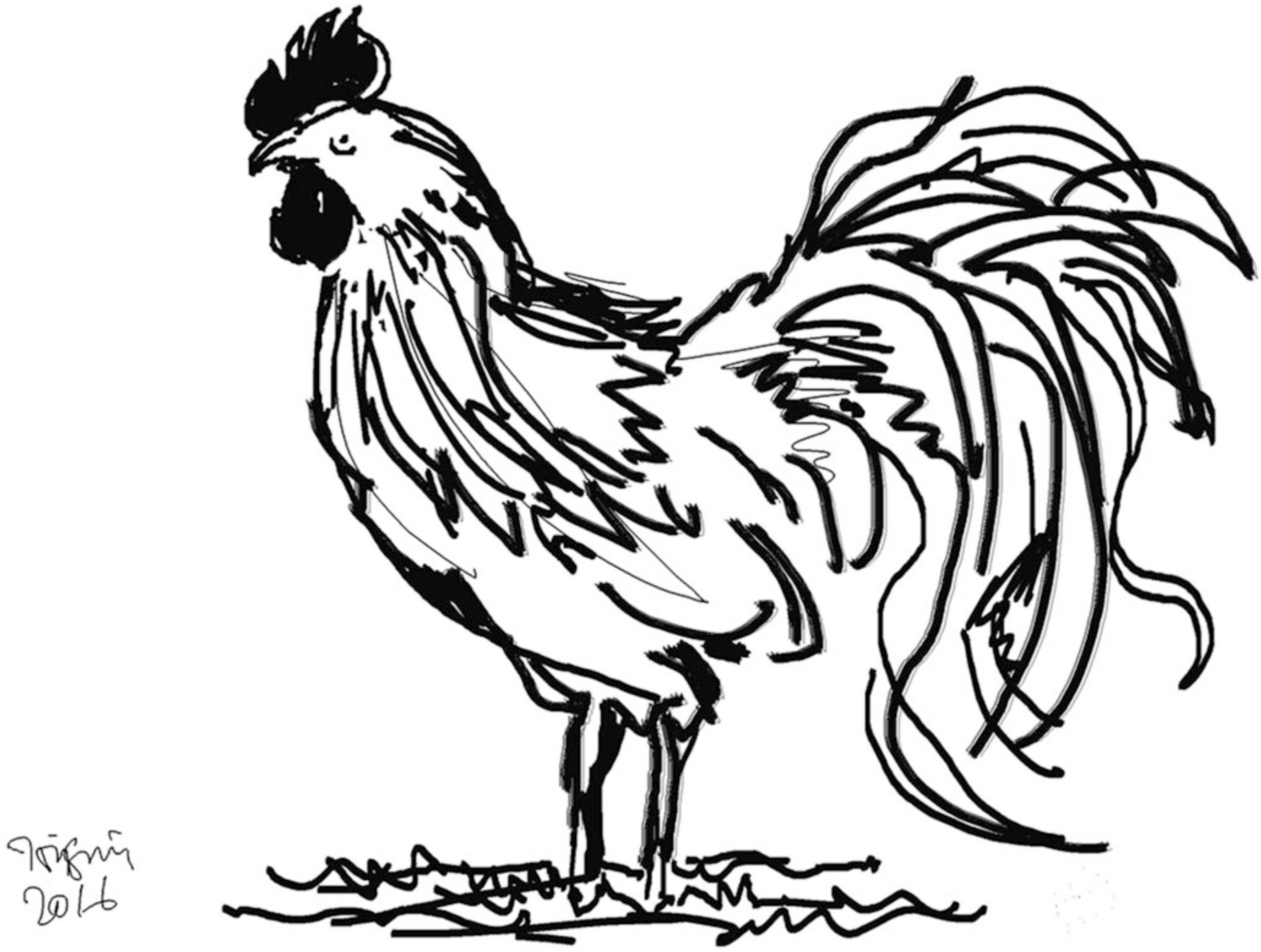
8 Multi Environmental Trials



Electronic supplementary material: The online version of this chapter (doi:10.1007/978-3-319-55177-7_8) contains supplementary material, which is available to authorized users.

F. Isik et al., *Genetic Data Analysis for Plant and Animal Breeding*, DOI 10.1007/978-3-319-55177-7_8

Abstract

Most field tests for plant breeding are replicated across different environments to measure the performance of breeding stocks across a range of environmental conditions to which a cultivar might be exposed. Multi-environment trials provide information about the adaptability of genotypes to specific environments or to sets of environments. The variance-covariance structures introduced in preceding chapters can be used to model genotype-by-environment interactions in multi-environmental trials. The number of parameters required to fit fully specified multi-environment trial models increases faster than the number of environments, so more parsimonious models are preferred when the number of environments is large. In this chapter we compare an unstructured matrix that involves separate parameters for genetic variance within each environment and for genetic correlations between each pair of environments to more parsimonious models, such as factor analytic structures, which require fewer parameters. Factor analytical structures can often efficiently capture the genotype-by-environment patterns without requiring extraordinary model complexity.

Introduction

In multi-environmental trials (MET) a set of genotypes or families are raised in a number of environments. The objectives are to compare genotypes across a range of environments and identify those that are generally adaptable across the testing environments, or to identify superior genotypes for subsets of the testing environments. If broad adaptation is not possible, then the breeder may instead prioritize selecting different genotypes with good performance in subsets of the environments. Proper analysis of MET data can reveal not only which genotypes are 'best' overall or in subsets of environments, but also can reveal the relationships among environments in terms of the genotype by environment (GxE) interaction patterns. This information can be used to improve the efficiency of breeding programs by identifying highly correlated clusters of environments that may represent oversampling of similar environments.

In addition to using METs to estimate GxE interactions, METs can serve a practical purpose in reducing the risk of losing the genetic materials due to environmental catastrophes. In many cases, breeders test a subset of material that is available in a given year and establish new field trails as new material becomes available. A subset of the genetic material (such as 'check' varieties) is used across multiple years to establish connections between testing series (years). A series of field trials established over time are also METs.

For yield and growth traits, large differences are often observed among environments. This occurs because of variation in soil fertility, precipitation, temperatures, and pathogen pressure. In perennial species, additional variation may be introduced because the ages of tests may differ among environments. For example, different growth rates may cause significant GxE interaction due to differences in the magnitude of genotypic variances across sites, even if genotype ranks do not change across environments (Cockerham 1963; Cooper and DeLacy 1994). This is a form of GxE interaction that does not hinder breeding gains, but is simply caused by the scale effect.

Ignoring heterogeneity in the variances can introduce bias in the predictions of breeding values and estimates of genetic variances, particularly if the breeding units are not replicated across all environments (Hill 1984). Accounting for heterogeneity in the data improves the accuracy of evaluations. In crop trials and in forest tree field tests, which may be balanced across sites within a year, accounting for heterogeneity of error variances in the mixed model can improve genotype predictions by giving more weight to information from environments with lower error variances.

Historically, such problems were not easy to handle with ordinary least squares ANOVA, but the flexibility of mixed models permits fitting complex multi-environment models that account for differences between residual as well as genotypic variances among sites. Further, mixed models approaches allow modelling of the pairwise genetic correlations among environments that provide a more realistic treatment than assuming that all pairs of environments have a common correlation, as was done in traditional ANOVA.

MET: General Approach and Considerations

Depending on the number of sites, we can perform one-stage or two-stage MET analysis. One stage is preferable but may not be feasible if there are large numbers of trials. Two-stage analysis proceeds as follows:

- Analyse data for each environment separately to check the data quality and estimate means and variances. We recommend this step even for one-stage analysis.
- If field position coordinates of plots are available, select the optimal spatial model for each site and predict site-specific genotype values for varieties
- Save predictions and their standard errors in a file
- Conduct a combined analysis across the sites based on the site-specific predictions. In combined analysis, some or most of the variances can be fixed to help with model convergence.
- The second stage can be weighted by the inverse of the variances of predictions of values from the first stage (Welham et al. 2010).

With increased computing power, one-stage analysis has become more feasible for large data sets. ASReml facilitates fitting different models for within-environment non-genetic effects and variation for different sites in the multi-environment single stage analysis. Differences among sites can include: (1) different field designs and covariates, (2) different spatial models within sites, and (3) heterogeneous variances across sites.

Modelling genetic correlations between each pair of sites using an unstructured (US) covariance matrix is feasible if there are a few sites and many entries. For s environments, the US covariance model requires s within-environment genetic variances and $s(s-1)/2$ pairwise environment covariances. If there are many sites, the number of parameters to estimate becomes very large. In such cases, factor analytic (FA or XFA) models are often more appropriate for modelling complex GxE patterns because they are more parsimonious, involving fewer parameter estimates. ASReml also allows fixing some variances and correlations that are at the boundary of theoretically allowable values to help models converge. For a given MET data set, we can consider a hierarchy of models of increasingly complex variance-covariance structures for both residuals (the **R** matrix) and genotype-environment effects (the **G** matrix). Like spatial analysis of field trials, a major focus of MET analysis is on selecting the best fitting model (while avoiding over fitting) to account for heterogeneity and predict the breeding values of genetic entries with high confidence.

Typical **R** structures that can be tested in MET analyses include:

- **IDV** *structure*: one common error variance for all environments
- **DIAG** *structure*: heterogeneous error variances across sites
- **AR1** *structure*: heterogeneous spatially correlated R structure: each environment has unique error variance and two-dimensional spatial error correlation pattern.

Commonly used **G** structures for MET include:

- **DIAG** *structure*: each environment has a unique genetic variance, but there are no correlations between environments (1 parameter for **G**)
- **CORUV** *structure*: constant genetic correlation between environments and genetic variance within environments (2 parameters for **G**). We show below that this is the traditional ANOVA structure for multi-environment models. This structure is also called "compound symmetry."
- **CORUH** *structure*: constant genetic correlations between pairs of environments but heterogeneous genetic variances within environments (with $s + 1$ variance parameters). If there are $s = 10$ environments, then $s + 1 = 11$ variance parameters are needed.
- **US** *structure*: unstructured covariance and heterogeneous variance. Each environment has a unique genetic variance and each pair of environments has a unique covariance, with $s(s + 1)/2$ variance parameters. For 10 environments, $10(11)/2 = \mathbf{55}$ variance parameters are needed.
- **CORGH** = **US** *structure*: This is also a fully heterogeneous genetic correlation and variance structure, so is equivalent to the US structure, but it is parameterized in terms of correlations instead of covariances between environments.
- **FA*n*** and **XFA*n*** *structures*: Factor analytic and extended factor analytic models that model heterogeneous within-environment genetic variances and unique pairwise correlations between environments, but the correlations are constrained to capture only the first n multivariate factors in the data. This requires $s(k+1)-k(k-1)/2$ parameters, where k is the number of factors modelled. For ten environments, an FA1 model requires **20** parameter estimates. This is a large reduction compared to US and CORGH structures.

Statistical Models

The classical ANOVA model for a cross-classified design of m genotypes evaluated at s environments with b complete blocks at each site is:

$$\boldsymbol{Y_{ijkl} = \mu + E_i + G_j + GE_{ij} + B(E)_{ik} + \varepsilon_{ijkl}} \tag{8.1}$$

where μ is the overall mean, E_i is the fixed effect of environment i; G_j is the random effect of genotype $j, G_j \sim N\left(0, \sigma_G^2\right)$; GE_{ij} is the random interaction between genotype j and environment i, $GE_{ij} \sim N\left(0, \sigma_{GE}^2\right)$; $B(E)_{ik}$ is the random effect of block k nested in environment $i, B(E)_{ik} \sim N\left(0, \sigma_B^2\right)$; ε_{ijkl} is the residual error associated with the experimental unit l of genotype j in k-th block of environment i, $\varepsilon_{ijkl} \sim N\left(0, \sigma_\varepsilon^2\right)$.

From the analysis of variance, we can estimate the variance components and compute the means of genotypes at specific sites and across all environments. Importantly, this model assumes that there are no correlations between different factors in the model. Based on that assumption, the covariance between the values of a genotype at two environments i and $i^{'}$ is:

$$Cov\left(Y_{ij.}, Y_{i'j.}\right) = Cov\left(G_j, G_j\right) + Cov\left(GE_{ij}, GE_{i'j}\right) = \sigma_G^2 + 0 \tag{8.2}$$

Where $Y_{ij.}$ and $Y_{i'j.}$ are values of a genotype j at two environments; G_j is the genetic value of genotype j, which is the same at the two environments; GE_{ij} and $GE_{i'j}$ are interaction effects of genotype j with the environments. By definition, the covariance of the genotype j effect with itself is the variance of genotype effects. The covariance between interaction effects of genotype j with two different environments is zero. So, the covariance of genotype j at two environments is the variance of genotypes (σ_G^2). The variance of true genotypic values within an environment (measured without error) is:

$$Var\left(Y_{ij.}\right) = Var\left(G_j\right) + Var\left(GE_{ij}\right) = \sigma_G^2 + \sigma_{GE}^2 \tag{8.3}$$

So, the correlation between true values of one genotype at any two sites is

$$r\left(Y_{ij},\ Y_{i'j}\right) = \frac{Cov\left(Y_{ij},\ Y_{i'j}\right)}{\sqrt{Var\left(Y_{ij}\right)Var\left(Y_{i'j}\right)}} = \frac{\sigma_G^2}{\sqrt{\left(\sigma_G^2 + \sigma_{GE}^2\right)\left(\sigma_G^2 + \sigma_{GE}^2\right)}} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE}^2} \tag{8.4}$$

The ratio $\sigma_G^2/\left(\sigma_G^2 + \sigma_{GE}^2\right)$ is sometimes called a 'type B genetic correlation'. Typically, type B genetic correlations refer to correspondence in performance of family means at different environments (Yamada 1962). The ratio is bounded as $0 \leq r_B \leq 1$. A value of $r_B = 0$ indicates no correspondence between performance of a genotype in different environments, whereas $r_B = 1$ suggests perfect correspondence between performance of genotypes in different environments (Burdon 1977). If we analyze two sites separately and estimate breeding values of genotypes tested at these two sites, the product-moment correlation between breeding values would be similar to r_B.

Thus, this model assumes that the genotypic variances expressed within all environments are equal: $\sigma_{G1}^2 = \sigma_{G2}^2 = \sigma_{G3}^2, \ldots, = \sigma_{Gs}^2$ and that the correlation of genotypic values between environments is the same for all pairs of environments, $r_{12} = r_{13}\ ,\ \ldots\ ,\ = r_{(s\ -\ 1)s}$. The mixed model approach will allow us to relax these assumptions, but the way to do this may not be immediately obvious, as it combines the genotype and genotype-by-environment factors into a single compound model factor of genotype nested within environment: $G(E)_{ij}$. This formulation then allows us to specify the pattern of genotypic variances within environments and also the correlation structure for the effects of a common genotype across environments. We start by specifying the nested model as

$$Y_{ijkl} = \mu + E_i + G(E)_{ij} + B(E)_{ik} + \varepsilon_{ijkl} \tag{8.5}$$

The effects in the model are the same as in the cross-classified model given in Eq. 8.1, but we have combined G_j and GE_{ij} into a single factor, $G(E)_{ij}$. We can start with the assumption that the distribution of $G(E)_{ij}$ is identical and independently distributed (*iid*): $G(E)_{ij} \sim N\left(0, \sigma^2_{G(E)}\right)$. Under this assumption, the covariance between values of a common genotype at different environments is zero:

$$Cov\left(Y_{ij}, Y_{i'j}\right) = Cov\left[G(E)_{ij}, G(E)_{i'j}\right] = 0 \tag{8.6}$$

and the variance of true genotypic values within environments is due solely to genotype-by-environment interaction variances as they were defined in the cross-classified model:

$$Var\left(Y_{ij}\right) = Var\left[G(E)_{ij}\right] = \sigma^2_{GE} \tag{8.7}$$

Of course, the *independent, identical distribution* assumption is usually worse than the original cross-classified model we started with, but writing the model in this form and using mixed models analysis gives great flexibility to specify a range of alternate assumptions and model forms. For example, we can make the model equivalent to the cross-classified analysis by changing the variance-covariance structure of the compound *G(E)*$_{ij}$ effects so that they have a common variance within environments and a common covariance across environment pairs (three environments in this example):

$$Cov\left(Y_{ij}, Y_{i'j}\right) = Cov\left[G(E)_{ij} + G(E)_{i'j}\right] = \sigma^2_G$$

$$Var\left[G(E)_{ij}\right] = \begin{bmatrix} \sigma^2_G + \sigma^2_{G(E)} & \sigma^2_G & \sigma^2_G \\ \sigma^2_G & \sigma^2_G + \sigma^2_{G(E)} & \sigma^2_G \\ \sigma^2_G & \sigma^2_G & \sigma^2_G + \sigma^2_{G(E)} \end{bmatrix} \otimes \mathbf{I}_m = \sigma^2_G \tag{8.8}$$

where $\mathbf{I}_m$ is the identity matrix with $m \times m$ dimensions for m genotypes. For example, with three environments, the variance-covariance matrix in Eq. 8.8 has dimension 3 × 3. By changing the structure of the 3 × 3 matrix in this Kronecker product, we can then allow for genotypic variances and covariances to vary among environments and pairs of environments, respectively. For example, at the other extreme of model complexity, we can allow each environment to have its own genetic variance and each pair of environments to have their own covariance. This is the unstructured (US) covariance model for genotype within environment effects and it involves six unique parameters:

$$Var\left[G(E)_{ij}\right] = \begin{bmatrix} \sigma^2_{G(E1)} & \sigma^2_{G21} & \sigma^2_{G31} \\ \sigma^2_{G12} & \sigma^2_{G(E2)} & \sigma^2_{G32} \\ \sigma^2_{G13} & \sigma^2_{G23} & \sigma^2_{G(E3)} \end{bmatrix} \otimes \mathbf{I}_m = \sigma^2_G \tag{8.9}$$

The US covariance formulation of the **G** matrix for METs involving large numbers of genotypes and environments may often fail to converge. For example, the unstructured **G** matrix in an experiment involving 50 environments requires estimation of 1275 parameters. Clearly, estimation of such a large number of parameters can be computationally prohibitive.

Factor analytic (FA) covariance structures for METs offer a more parsimonious approach to capture the complexity of covariances among many environments while limiting the number of parameters that require estimation (Smith et al. 2001, 2005; Thompson et al. 2003). For s trials, the number of parameters to be estimated for the US model is $p = s(s + 1)/2$, whereas for FA models it is $s(k + 1) - k(k - 1)/2$, where k is the number of factors (Thompson et al. 2003). The reduction in parameters requiring estimation can be noted for the case of 50 environments and $k = 1$ factor, for which only 100 parameters are estimated compared to the 1275 required for the unstructured model.

The US and compound symmetry models can be formulated as specific cases of the FA model. For example, if we fit the maximum of $k = s - 1$ factors, we recapitulate the US model with $s(s + 1)/2$ parameters. At the other extreme, we can create the compound symmetry model in this framework by fitting $k = 1$ factor and forcing the site loadings (explained below) to be equal, requiring only two parameters, one factor to generate the correlation between environments and one variance component (Cullis et al. 2014; Meyer 2009).

If the vector of genetic effects nested within sites is written as $\mathbf{u_g}$, we can conceive of these effects being arranged as a matrix of effects with *m* rows (for *m* genotypes) and *s* columns (for *s* environments). Conceptually, then, this matrix of effects can be subjected to factor analysis, in which the patterns of genotype response across environments are modelled as interactions between genotype effects and one or a small number of factors that underlie the environmental influences on genotype-within-environment phenotypes. FA models can be interpreted as random regression models of genotype and GE effects on *k* unknown environmental covariates, in which each genotype has its own slope (genotypic scores) but a common intercept (Crossa et al. 2006). The slopes measure the sensitivity of genotypes to hypothetical environmental factors represented in the model by the numerical 'loadings' for each site in each factor (Piepho et al. 2007; Smith et al. 2005). In this model, the genotypic effect for genotype *j* in site *i* (u_{gij}) is a sum of *k* multiplicative terms (Cullis et al. 2014; Smith et al. 2002):

$$u_{gij} = \lambda_{1i} f_{1j} + \lambda_{2i} f_{2j} + \ldots + \lambda_{ki} f_{kj} + \delta_{ij} \tag{8.10}$$

The terms in the multiplicative model include λ_{1i}, the **loading** for environment *i* on the first factor; f_{1j}, the genetic effect (**score**) of genotype *j* on the first factor; λ_{ki}, the loading for environment *i* on factor *k; f_{kj}*, the score of genotype *j* on factor *k*, and δ_{ij} is the deviation of the observed genetic effect of genotype *j* in environment *i* from its predicted value based on the multiplicative factor model fit. Factor analysis is related to principal components analysis but whereas principal components decomposition of the matrix of GE effects would identify eigenvectors based on their ability to account for the variation within and covariance between environments, the FA model identifies factors that maximally explain the covariance among environments and introduces an additional unique variance to capture any additional variation within each environment.

The FA models are named based on the number of the *k* factors (multiplicative terms) included in the model, e.g., FA1, FA2, and FA*k*. Our hope is to identify a model that can accurately describe the observed variance-covariance relationships among and within environments with as few factors as possible.

For a given number of factors *k* selected, the covariance between a genotype's performance in different environments is estimated as (Smith et al. 2002):

$$Cov\left(Y_{ij}, Y_{i'j}\right) = \sum_{f=1}^{k} \lambda_{fi}\lambda_{fi'} \tag{8.11}$$

Notice that this generates a unique covariance for each pair of environments if loadings differ among the environments. The variance of genotypic effects within an environment is estimated as:

$$Var\left(Y_{ij}\right) = \sum_{f=1}^{k} \lambda_{fi}^{2} + \sum_{j=1}^{m} \frac{Var\left(\delta_{ij}^{2}\right)}{m} \tag{8.12}$$

The second piece of this expected variance is the average site-specific variance over all *m* genotypes within environment *i*. This is the within-site variance that is not accounted for by the factor loadings, and will be designated Ψ_{gi} (Smith et al. 2002).

$$Var\left(Y_{ij}\right) = \sum_{m=1}^{k} \lambda_{mi}^{2} + \Psi_{gi} \tag{8.13}$$

Writing the vector of genotypic effects within environments in the $m \times s$ matrix form, we have:

$$\mathbf{u}_g = (\mathbf{\Lambda} \otimes \mathbf{I}_m)\boldsymbol{f} + \boldsymbol{\delta} \tag{8.14}$$

Where $\mathbf{I}_m$ is the identity matrix with dimensions $m \times m$, $\mathbf{\Lambda}$ is the matrix of environment loadings (with dimension $s \times k),\boldsymbol{f}$ is the vector of genotypic scores with dimensions $mk \times 1$, and $\boldsymbol{\delta}$ is a vector of residual genetic effects (with dimensions $ms \times 1$). If the genotypic effects are additive breeding values with an additive relationship matrix $\mathbf{A}$, then the variances of $\boldsymbol{f}$ and $\boldsymbol{\delta}$ are:

$$Var(\boldsymbol{f}) = \mathbf{A} \otimes \mathbf{I}_k \tag{8.15}$$

$$Var(\boldsymbol{\delta}) = \mathbf{A} \otimes \mathbf{\Psi} \tag{8.16}$$

where $\mathbf{I}_k$ is a $k \times k$ identity matrix and $\mathbf{\Psi}$ is an $s \times s$ diagonal matrix with site-specific genetic variances (ψ_s) on the diagonal and zero covariance between sites. $\mathbf{A}$ can be replaced with $\mathbf{I}_m$ if relationships are unknown and families are assumed independent, or with some other relationship matrix, such as the realized relationship matrices described in Chap. 11.

The variance of additive genotypic effects across all trials is:

$$Var(\mathbf{u}_g) = \mathbf{G}_g = \left(\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}\right) \otimes \mathbf{A} \tag{8.17}$$

Typically, the model fitting process starts by fitting an FA1 model and proceeds to fit more complex ($k > 1$) models. Since the models are nested we can use likelihood ratio tests (LRT), Akaike Information Criterion (AIC), or Bayesian Information Criterion (BIC) to select models, although at some point model convergence may hinder fitting more complex models and we can stop. Smith et al. (2014) suggested that AIC and LRT might select models that are too complex (overfit), whereas BIC which penalizes model complexity more, might select underfit models that miss some important signal in the data. They suggest measuring goodness-of-fit for each model based on both the percent variance explained by k factors at within each individual environment (V_i) and averaged across environments ($\bar{\mathrm{V}}$) as follows

$$V_i = 100\frac{\sum_{r=1}^{k}\lambda_{ri}^2}{\sum_{r=1}^{k}\lambda_{ri}^2 + \psi_i^2} \tag{8.18}$$

$$\bar{V} = 100\frac{tr\left(\Lambda\Lambda^T\right)}{tr\left(\Lambda\Lambda^T + \mathbf{\Psi}\right)} \tag{8.19}$$

where $tr()$ is the trace of the matrix (sum of diagonal elements) (Smith et al. 2014). The first factor accounts for as much of the covariances of genotype performances among environments as possible; subsequent factors are independent of previous factors and explain consecutively less covariance. Smith et al. (2015) recommend a model where the proportion of variation within most environments is high and few environments have low variance explained. These metrics are useful diagnostics, but unfortunately, they do not provide a model selection criterion. The choice of the number of factors to fit remains complicated; ideally a few factors can capture most of the patterns in the observed data, which is ideal for reducing the number of parameters.

Formulation of FA models in ASReml

In ASReml, FA models are specified in a covariance form, correlation form, or in an extended factor analytic (XFA*k*) form (Gilmour et al. 2014). In the ***covariance*** formulation of FA models, the variance is given as the direct product of an FA covariance matrix for sites (environments) and a genotype effect correlation matrix (which could be IDV or a numerator or other relationship matrix for genotype effects). The FA covariance structure for sites is parameterized as $\mathbf{\Lambda\Lambda}^{\boldsymbol{T}} + \mathbf{\Psi}$, where $\mathbf{\Lambda}$ is the matrix of loadings on the covariance scale. As an example, the covariance matrices for FA1 model with m unrelated genotypes tested at four sites would be:

k = 1 factor

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{14} \end{bmatrix}, \mathbf{\Psi} = \begin{bmatrix} \Psi_1 & 0 & 0 & 0 \\ 0 & \Psi_2 & 0 & 0 \\ 0 & 0 & \Psi_3 & 0 \\ 0 & 0 & 0 & \Psi_4 \end{bmatrix},$$

$$\mathbf{G}_{\boldsymbol{g}} = \left[\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}\right] \otimes \mathbf{I}_{\boldsymbol{m}} = \begin{bmatrix} \lambda_{11}^2 + \Psi_1 & \lambda_{11}\lambda_{12} & \lambda_{11}\lambda_{13} & \lambda_{11}\lambda_{14} \\ \lambda_{11}\lambda_{12} & \lambda_{12}^2 + \Psi_2 & \lambda_{12}\lambda_{13} & \lambda_{12}\lambda_{14} \\ \lambda_{11}\lambda_{13} & \lambda_{12}\lambda_{13} & \lambda_{13}^2 + \Psi_3 & \lambda_{13}\lambda_{14} \\ \lambda_{11}\lambda_{14} & \lambda_{12}\lambda_{14} & \lambda_{13}\lambda_{14} & \lambda_{14}^2 + \Psi_4 \end{bmatrix} \otimes \mathbf{I}_{\boldsymbol{m}}$$

The covariance matrices for FA2 model with m unrelated genotypes tested at four sites would be:

$k = 2$ *factors*

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{21} \\ \lambda_{12} & \lambda_{22} \\ \lambda_{13} & \lambda_{23} \\ \lambda_{14} & \lambda_{24} \end{bmatrix},$$

$$\mathbf{G}_g = \begin{bmatrix} \lambda_{11}^2 + \lambda_{21}^2 + \Psi_1 & \lambda_{11}\lambda_{12} + \lambda_{21}\lambda_{22} & \lambda_{11}\lambda_{13} + \lambda_{21}\lambda_{23} & \lambda_{11}\lambda_{14} + \lambda_{21}\lambda_{24} \\ \lambda_{11}\lambda_{12} + \lambda_{21}\lambda_{22} & \lambda_{12}^2 + \lambda_{22}^2 + \Psi_2 & \lambda_{12}\lambda_{13} + \lambda_{22}\lambda_{23} & \lambda_{12}\lambda_{14} + \lambda_{22}\lambda_{24} \\ \lambda_{11}\lambda_{13} + \lambda_{21}\lambda_{23} & \lambda_{12}\lambda_{13} + \lambda_{22}\lambda_{23} & \lambda_{13}^2 + \lambda_{23}^2 + \Psi_3 & \lambda_{13}\lambda_{14} + \lambda_{23}\lambda_{24} \\ \lambda_{11}\lambda_{14} + \lambda_{21}\lambda_{24} & \lambda_{12}\lambda_{14} + \lambda_{22}\lambda_{24} & \lambda_{13}\lambda_{14} + +\lambda_{23}\lambda_{24} & \lambda_{14}^2 + \lambda_{24}^2 + \Psi_4 \end{bmatrix} \otimes \mathbf{I}_m$$

In the ***correlation*** parameterization of FA models, the factor loadings are scaled by the genetic variances within sites. The matrix of loadings is now referred to as **F** and is analogous to the **Λ** matrix in the covariance form. For example, for an FA1 model on the correlation scale:

$$\boldsymbol{F} = \begin{bmatrix} f_{11} \\ f_{12} \\ f_{13} \\ f_{14} \end{bmatrix} = \begin{bmatrix} \dfrac{\lambda_{11}}{\sqrt{\lambda_{11}^2 + \Psi_1}} \\ \dfrac{\lambda_{12}}{\sqrt{\lambda_{12}^2 + \Psi_2}} \\ \dfrac{\lambda_{13}}{\sqrt{\lambda_{13}^2 + \Psi_3}} \\ \dfrac{\lambda_{12}}{\sqrt{\lambda_{14}^2 + \Psi_4}} \end{bmatrix}$$

The off-diagonal elements of the product $\mathbf{FF}^T$ are the correlations between environments. However, $\mathbf{FF}^T$ is not a correlation matrix because its diagonal elements are not equal to 1. Therefore, we create a correlation matrix **C** by adding to the diagonal elements of $\mathbf{FF}^T$ to make them 1:

$\mathbf{C} = \mathbf{FF}^T + \mathbf{E}$, where **E** is a diagonal matrix defined as $\mathbf{E} = \mathrm{diag}(1 - \mathbf{F}^2)$.

We can then generate the covariance matrix for sites as **DCD**, where $\boldsymbol{D}$ is an $s \times s$ diagonal matrix whose elements are square roots of the genetic variance within each site, i.e. $\boldsymbol{D}_{11} = \sqrt{\lambda_{11}^2 + \Psi_1}$ for an FA1 model. Then the covariance structure for lines within sites is:

$$\mathbf{G}_g = [\boldsymbol{DCD}] \otimes \mathbf{I}_m$$

Notice that **DF** in the FA correlation model is equal to **Λ** in the FA covariance formulation. Similarly, **DED** in the FA correlation formulation is equal to **Ψ** in the FA covariance formulation (Gilmour et al. 2014).

The *covariance* and *correlation* formulations of model parameterizations can have convergence problems and produce zero or even negative site-specific variances. This can occur when the factors alone ($\mathbf{\Lambda\Lambda}^T$ or $\mathbf{FF}^T$) explain all of the variance within a site or predict more variance than is actually observed, such that one or more elements of **Ψ** or $\boldsymbol{E}$ are zero or negative. This situation is referred to as a *Heywood* case (Smith et al. 2001). ***Extended factor analytical*** (XFAk) models were developed to avoid convergence problems related to *Heywood* cases and also to increase computational efficiency (Meyer 2009; Thompson et al. 2003). XFA models have the same parameterization as FA covariance models $\mathbf{\Lambda\Lambda}^T + \mathbf{\Psi}$, but the algorithm used to fit the model is different. The common factors ($\lambda_{ri}f_{rj}$) are fit separately from the specific factors ($\delta_{g_{ij}}$), which leads to greater sparsity in the mixed model equations; furthermore, if a site-specific variance is zero, the $\delta_{g_{ij}}$ effects at that site are set to zero without hindering convergence for estimating the other model effects (Meyer 2009; Thompson et al.

2003). In ASReml syntax, the parameters in the *covariance* and *correlation* models are specified in the order of loadings **Λ** or **F** followed by specific variances, **Ψ** or **E**. In contrast, in XFA*k* models, the specific variances, **Ψ,** are specified first, followed by the loadings, **Λ**.

Example: Analysis of Pine Polymix MET Data

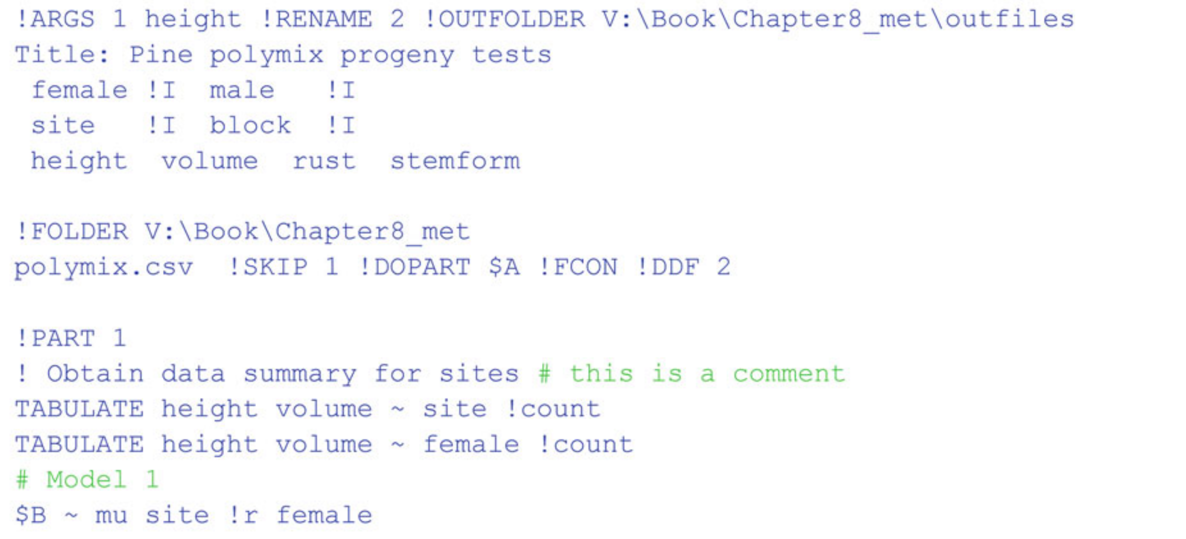
Polymix mating involves pollinating a set of individuals using bulked pollen from another set of individuals to reduce the cost of breeding. The goal is to predict breeding values of females for half-sib family selection. The Cooperative Tree Improvement Program at North Carolina State University used polymix breeding in the third cycle of loblolly pine (*Pinus taeda* L.) selection to predict the general performance of female parents (McKeand and Bridgwater 1998). In one of the test series, 70 individuals were mated with bulked pollen collected from another set of 40 individuals. Progeny from crossing were considered half-sibs with known mother and different fathers. A randomized complete block design was used with 20 blocks. Each female parent had one progeny in each block, for a total of 20 progeny at a site at the time of the planting. The experiment was replicated at 12 sites in the southeastern US. Height of tree, stem volume, fusiform rust disease incidence (present = 1, absent = 0) and stem straightness (1–6, 1 being the most strait) were assessed at age 6 years. A subset of the data is given below (*polymix.csv*).

```
female male site block height volume rust stemform
    16    0  101    18   25.5   0.76    0        1
    16    0  101     5   21.5   1.09    0        4
    16    0  101     6   24.5   1.19    0        4
    16    0  101     3   29.0   2.15    0        5
    16    0  101    17   32.0   2.85    0        2
    16    0  101    16   27.0   1.64    0        3
```

Summarize Data for Each Site

This section allows us to check that our program is reading the data correctly and also provides summary statistics for each site. The only terms included in the linear model are fixed intercept and site effects and a random female effect.

Code example 8.1
Analysis of pine polymix data (see Code 8-1**_MET.as for more details)**



```
!ARGS 1 height !RENAME 2 !OUTFOLDER V:\Book\Chapter8_met\outfiles
Title: Pine polymix progeny tests
 female !I  male   !I
 site   !I  block  !I
 height  volume  rust  stemform

!FOLDER V:\Book\Chapter8_met
polymix.csv  !SKIP 1 !DOPART $A !FCON !DDF 2

!PART 1
! Obtain data summary for sites # this is a comment
TABULATE height volume ~ site !count
TABULATE height volume ~ female !count
# Model 1
$B ~ mu site !r female
```

- The line starting with the exclamation point and space writes the text that follows to the primary output file (.asr). This is a way to include comments preceding the output.

- We requested tabulation (TABULATE) of height for each site. This will generate a file with a *.tab* extension including the mean, standard deviation, minimum, and maximum of plant height measures as well as the total number of observations for each site. The second TABULATE statement generates summary statistics for female parents.

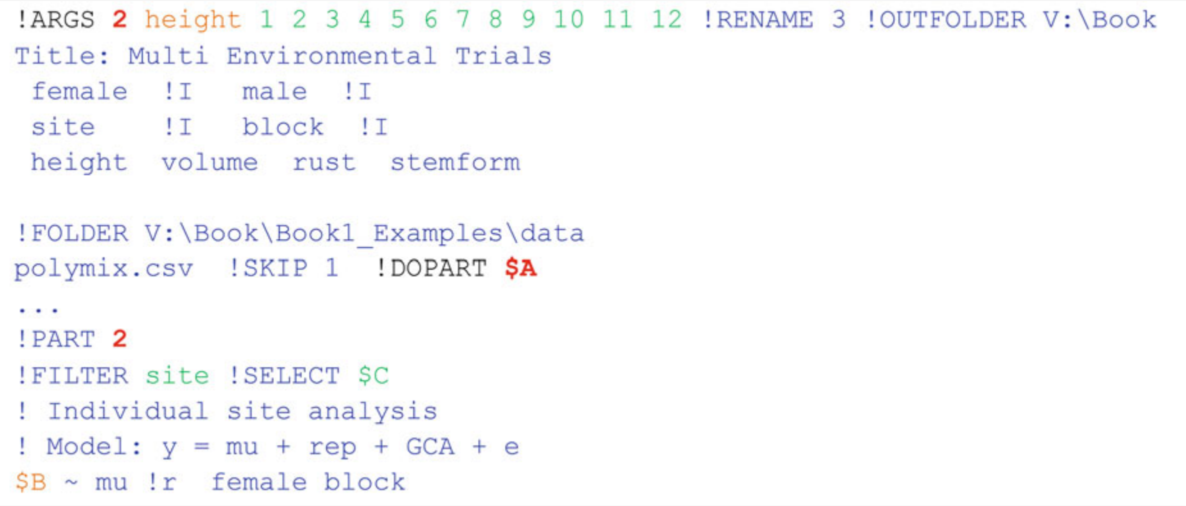
The output from TABULATE (*Code 8-1_MET1_height.tab*) includes descriptive statistics for sites. The range of site means for height growth is 21.8–29.7. The number of observations per site ranged from 1125 to 1372. The approximate F-tests in the primary output file *.asr* are given below.

```
Wald F statistics
  Source of Variation   NumDF    DenDF_con F-inc     F-con M P-con
9 mu                        1      69.4 79076.11  79076.11 . <.001
3 site                     11   15313.2  1205.49   1205.49 A <.001
```

The large and significant F-value for site effect indicates that the variation among sites is significant. A common observation in multi environmental trials (MET) data is that when the mean values for a trait vary significantly among environments, often the error variances may also differ significantly among site. This is often simply a matter of scale, with larger variances associated with larger observed measurements. In such cases, the default assumption of homogeneous residuals at all sites ($\varepsilon \sim N(0,\sigma^2_e\mathbf{I}_n)$) may not hold.

Analyze Each Site Separately to Obtain Variances

The second step is to analyse each site separately to obtain site-specific error variances and genetic variances. The model for each site is: $Y_{ijk} = \mu + B_i + G_j + \varepsilon_{ijk}$, where B_i is the random block effect, G_j is the random female effect and ε_{ijk} is the random residual. Variance components from individual sites can be used as starting parameters when we run the combined MET analysis and attempt to fit heterogeneous error variances. One way to run the same model for different sites is to use '`!FILTER site !SELECT n`' in combination with `!ARGS`:



```
!ARGS 2 height 1 2 3 4 5 6 7 8 9 10 11 12 !RENAME 3 !OUTFOLDER V:\Book
Title: Multi Environmental Trials
 female  !I   male  !I
 site    !I   block  !I
 height  volume  rust  stemform

!FOLDER V:\Book\Book1_Examples\data
polymix.csv  !SKIP 1  !DOPART $A
...
!PART 2
!FILTER site !SELECT $C
! Individual site analysis
! Model: y = mu + rep + GCA + e
$B ~ mu !r  female block
```

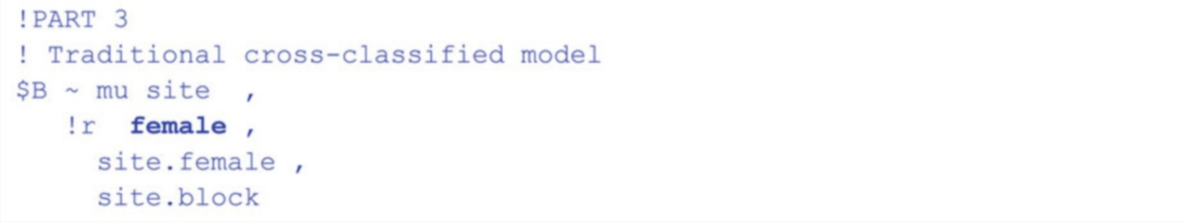
- The argument **`$A`** after naming the data file indicates the point at which the first argument ('2') will be substituted (the PART to analyse).
- The argument **`$B`** in the models indicates the point at which the second argument ('height') will be substituted (the trait to analyse).
- **`$C`** indicates the point at which the program will iteratively substitute the remaining arguments, one at a time ('1' through '12'). Here `$C` indicates the level of site to select when filtering the data set in the current iteration. `!FILTER` *v* `!`

SELECT *n* together are used to select data from a single site for analysis. The *v* is the number or name of a data field ('site' in this case) and *n* is the value of the field to be selected. It can be an integer (as in this example) or a character string in quotes. This is similar to using the BY statement in SAS procedures.

Different output files will be created for each site (file names will include three variable suffix values corresponding to PART, TRAIT, and SITE). In the output files we see large differences between sites for both genetic (range 0.45–1.48) and residual variances (3.18–11.41). Block differences at each site also explain considerable variation and should remain in subsequent multi-environment models. Heritability estimates had a range of 0.28–0.61. Now that we have a sense of the heterogeneity in the data, we will keep in mind that our final model should reflect this. Before we include such complexity in the combined model, however, we will start with the simplest model, a cross-classified ANOVA.

Model 3: Cross-Classified ANOVA

We can perform the combined analysis across environments using the traditional cross-classified genotype-environment model. The variance structures for random effects, including the residual, are scaled identical and independent (IID) variances. The linear model is $Y_{ijk} = \mu + S_i + SB_{ij} + F_k + SF_{ik} + \varepsilon_{ijk}$, where Y_{ijk} is the observation on a progeny of female *k* in block *j* at site *i*, S_i is the *i*-th site effect, SB_j is the random block effect nested within site, F_k is the random female effect, SF_{ik} is the random female by site interaction effect and ε_{ijkl} is the random residual associated with the data point. We can fit site as fixed effect since we have a balanced design in this case and we are not interested in making predictions or inferences about site effects or variances. Female is a random factor. Therefore the site-by-female interaction is random, even if site effect is considered fixed effect.



```
!PART 3
! Traditional cross-classified model
$B ~ mu site  ,
   !r  female ,
     site.female ,
     site.block
```

A small subset of the output from *.asr* file is given below.

OUTPUT 3

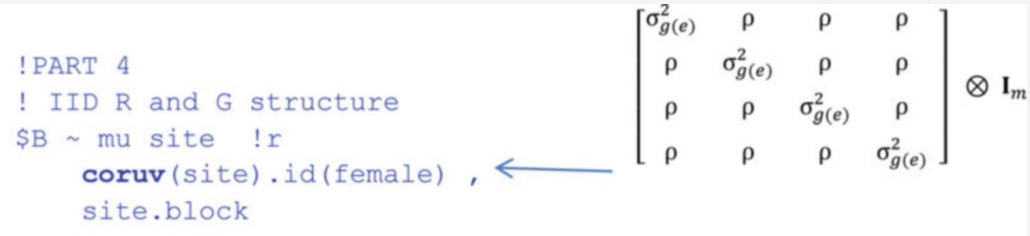
```
         7 LogL=-3323.03 S2= 7.1215 15379 df

- - - Results from analysis of height - - -
Akaike Information Criterion 46654.05 (assuming 4 parameters).
Bayesian Information Criterion 46684.61

Model_Term                        Sigma  Sigma/SE  %  C
female        IDV_V    70  0.563185      5.42   0  P
site.block    IDV_V   240  0.931300      9.53   0  P
site.female   IDV_V   840  0.174454      5.95   0  P
Residual      SCA_V 15391   7.12148     84.68   0  P
```

- All variance components seem to be significant since they are at least two times their standard errors (Sigma/SE column).

Model 4: Compound Symmetry

We can modify the factorial family-environment model to be a family nested within environment model as: $Y_{ijk} = \mu + S_i + SB_{ij} + SF_{ik} + \varepsilon_{ijk}$, where the terms are the same as above, except that by removing the family main effect, the family effects become nested within sites. We can recover a model equivalent to the cross-classified ANOVA model by fitting a *compound symmetry model* (`coruv` **G** structure) to the nested *site.female* effect. This fits a common genetic variance within sites and a common pairwise correlation between sites. We assume a uniform variance (coru**v**) for female within site effects and a uniform correlation (`coruv`) or covariance between pairs of sites. The model in ASReml is:

PART 4



```
!PART 4
! IID R and G structure
$B ~ mu site  !r
    coruv(site).id(female) ,
    site.block
```

$$\begin{bmatrix} \sigma^2_{g(e)} & \rho & \rho & \rho \\ \rho & \sigma^2_{g(e)} & \rho & \rho \\ \rho & \rho & \sigma^2_{g(e)} & \rho \\ \rho & \rho & \rho & \sigma^2_{g(e)} \end{bmatrix} \otimes \mathbf{I}_m$$

- The covariance structure shown to the right is for four sites only as an example.
- In the model there is no female main effect. It appears with site as a consolidated term (*site.female*).

OUTPUT 4

```
      8 LogL=-3323.03     S2=  7.1215      15379 df

         - - - Results from analysis of height - - -
  Akaike Information Criterion    46589.43 (assuming 15 parameters).
  Bayesian Information Criterion  46704.04

  Model_Term                            Sigma    Sigma/SE    % C
  site.id(block)          IDV_V  240   0.931300       9.53    0 P
  Residual               SCA_V 15391    7.12148      84.68    0 P
  coruv(site).id(female)  840 effects
  site                    COR_R    1   0.763497      16.76    0 P
  site                    COR_V    1   0.737639       6.88    0 P
```

- The parameters related to *site.female* effect are labeled with "*site* COR_R" for pairwise correlation between sites (identical for pairs of sites) or with "*site* COR_V" for the female within site variance component.

The relationship of COR_R and COR_V estimates from model 4 to the variance components from the cross-classified model (PART 3) may not be immediately obvious but they are indeed the same model. We have just changed how the model is parameterized. Notice that the residual LogL of models 3 and 4 are identical. Recall that the cross-classified ANOVA model produced a variance component estimate of 0.5632 for *female* effect and a variance component of 0.1744 for the *female.site* interaction effect. The sum of these two variance components from the ANOVA model (0.5632 + 0.1744) is equal to the variance component for the compound term *site.female* in the nested model, $Var\left(Y_{ij}\right) = Var\left(G(E)_{ij}\right) = \sigma^2_G + \sigma^2_{GE} = 0.737$. Further, the ratio of *female* to the sum of *female* and *site.female* variance components estimated from the ANOVA cross-classified model, 0.5632 / (0.5632 + 0.1744) = 0.76 is equal to the pairwise site correlation estimate from the nested CORUV model (COR_R = 0.76).

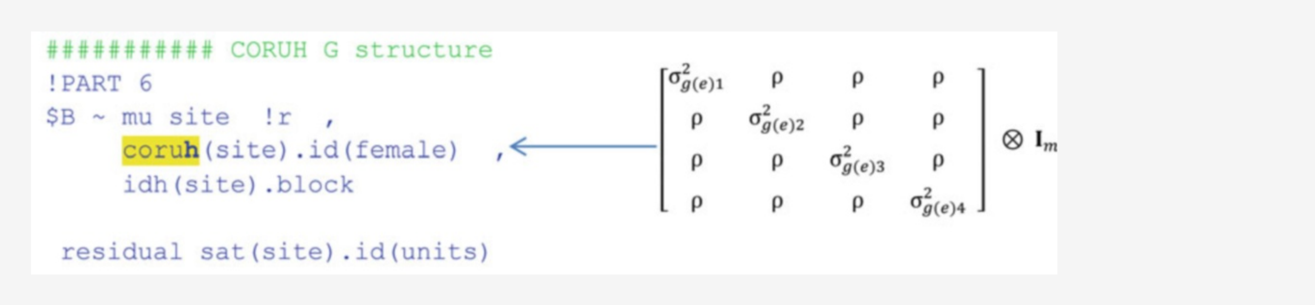
Model 5: Heterogeneous Residuals and Block Effects

In the models above we assumed that residuals and blocks nested within sites have scaled identity variance structures. However, we saw in part 2 that the models fit within each site separately resulted in widely different residual variances. Checking for heterogeneity in the residual variances across sites is a recommended practice. We can perform a formal test of the null hypothesis that the residual variances are uniform among sites by fitting the block diagonal residual structure `residual sat(site).id(units)`, which fits a separate residual variance for each site, and comparing the resulting log likelihood to model 3 or 4. The LogL for the heterogeneous R structure model was −2909 while it was LogL = −3323 for the homogeneous residual model (OUTPUT 4). The likelihood ratio test statistic would be 2(−2909 − (−3323)) = 828 with 11 degrees of freedom (1 residual variance versus 12). Clearly a chi-square value of 828 with 11 df is significant (the critical value of chi-square for 11 df is 19.67 at $p = 0.05$), so we can safely reject the null hypothesis of equal residual variances among sites. We can also test the assumption that the block within site variances are equal among environments by fitting a heterogenous block within site variance structure with the model term `idh(site).block` (or, equivalently, `at(site).block)`. The heterogeneous block within site variance model was also significantly better, so for the remaining examples in this chapter, we will use both heterogeneous residual and block variances across sites. Next, we will focus on fitting different **G** structures to model the variance-covariance relationships among family-within-site effects.

Model 6: CORUH G Structure

The compound symmetry structure, `coruv()` of genetic effects in models 4 and 5 assumes that the random genotype and genotype by environment interaction effects are constant. It involved only two genetic parameters; a variance and a correlation. This is an underfit model, as we shall see in the following models. We can relax a uniform G structure by allowing different genetic (*female*) variances at each site. This makes sense since there appeared to be large differences between sites for female variance components, with a range of 0.45 (site 12) to 1.48 (site 1) observed among the individual site models in part 2. Part 6 of our ASReml program fits a CORUH model:

PART 6



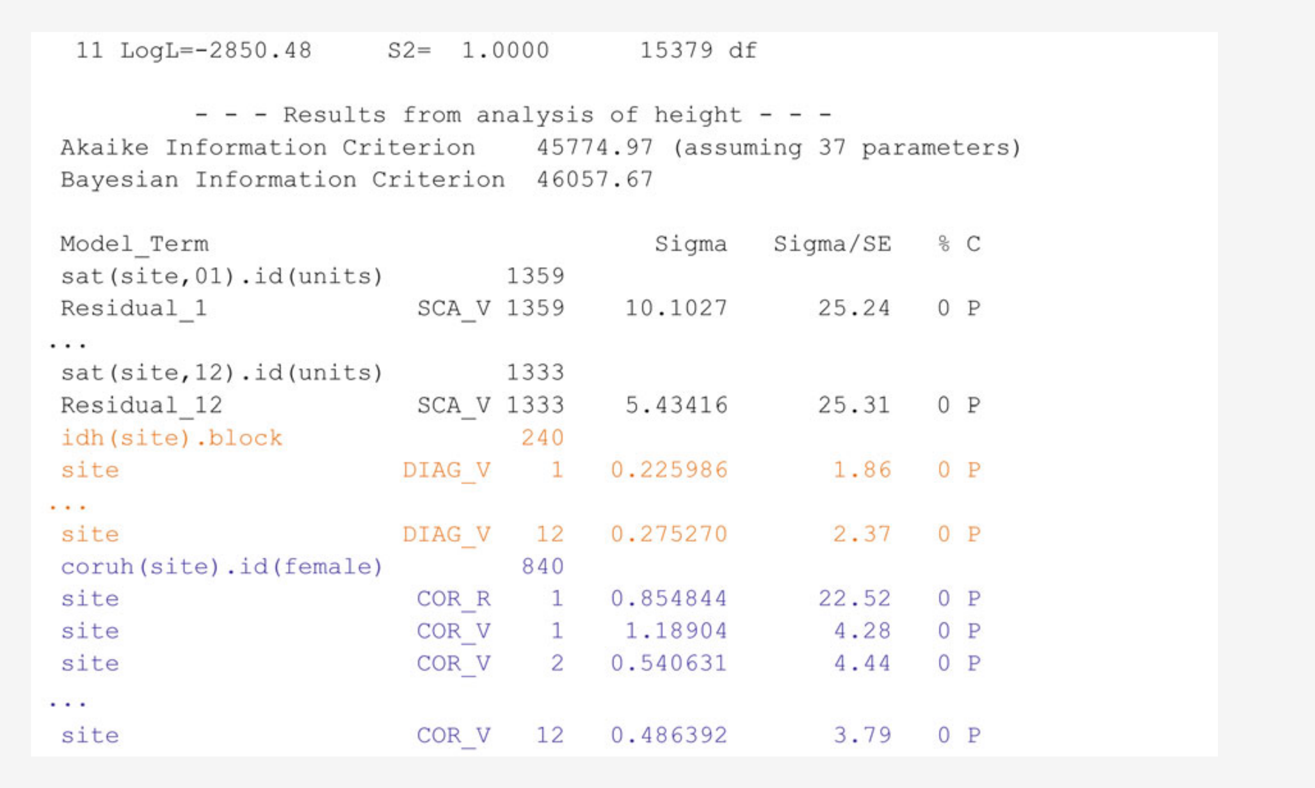
```
########### CORUH G structure
!PART 6
$B ~ mu site  !r  ,
     coruh(site).id(female)   ,
     idh(site).block

 residual sat(site).id(units)
```

$$\begin{bmatrix} \sigma^2_{g(e)1} & \rho & \rho & \rho \\ \rho & \sigma^2_{g(e)2} & \rho & \rho \\ \rho & \rho & \sigma^2_{g(e)3} & \rho \\ \rho & \rho & \rho & \sigma^2_{g(e)4} \end{bmatrix} \otimes \mathbf{I}_m$$

- The G structure for $F(E)_{ij}$ effects is a direct product of the $s \times s$ variance-covariance matrix for a common female's effects within and across sites (although we only show a matrix for four sites in the example above) and the identity matrix (assuming females are unrelated). The variance function `coru`**`h`**`()` fits a heterogeneous variance structure to female effects. The `coru` stands for uniform correlation, and **`h`** indicates heterogeneous variances.

A subset of the results is given below:

OUTPUT 6



```
 11 LogL=-2850.48     S2=  1.0000      15379 df

        - - - Results from analysis of height - - -
Akaike Information Criterion    45774.97 (assuming 37 parameters)
Bayesian Information Criterion  46057.67

Model_Term                              Sigma   Sigma/SE   % C
sat(site,01).id(units)        1359
Residual_1             SCA_V 1359    10.1027      25.24   0 P
...
sat(site,12).id(units)        1333
Residual_12            SCA_V 1333    5.43416      25.31   0 P
idh(site).block                240
site                  DIAG_V     1   0.225986      1.86   0 P
...
site                  DIAG_V    12   0.275270      2.37   0 P
coruh(site).id(female)         840
site                   COR_R     1   0.854844     22.52   0 P
site                   COR_V     1    1.18904      4.28   0 P
site                   COR_V     2   0.540631      4.44   0 P
...
site                   COR_V    12   0.486392      3.79   0 P
```

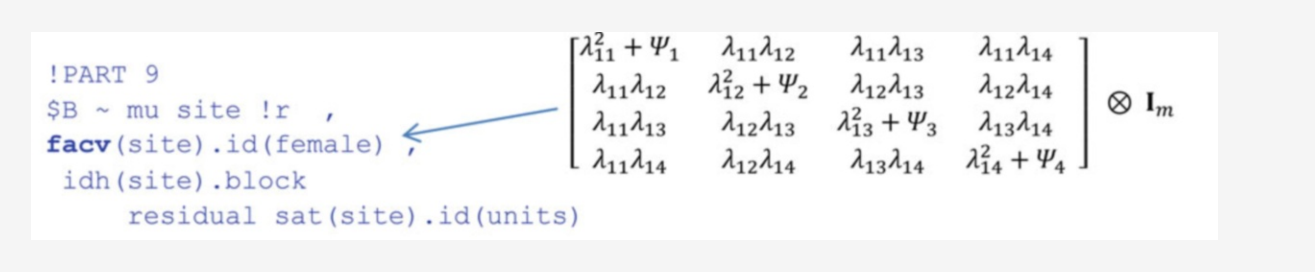
- The logL of model 6 is -2850.48. This is a substantial improvement over the modified model 5 that included heterogeneous block and residual variances (results not shown).

Models 7 and 8: US and CORGH Structures

In model 6 we relaxed the constant variance assumption and fit a heterogeneous variance structure for the female within site effect. However, the `coruh()` model may still be too restrictive because it assumes a constant genetic correlation between pairs of environments. The general form of the variance structure for female effects would have different variances at each environment and different correlations (or covariances) between pairs of environments. In other words, fitting the `us()` and `corgh()` structures with $p = \frac{s(s+1)}{2}$ parameters. As the number of environments increases, model convergence and reliability of parameters become an issue. Therefore these structures are not recommended for multi environmental models with large numbers of environments (Smith et al. 2005). In this example, the number of parameter estimates for the female within site effect for these models is 78. The ASReml code to fit these models are included as models 7 and 8 in the example code file “Code 8-1_MET.as”; we were able to attain convergence only for model 8 (the CORGH model), which had a log likelihood of -2804.06, Akaike Information Criterion of 45812.12, and Bayesian Information Criterion of 46591.48.

Model 9: FA1 Covariance Structure

In PART 9 we fit the FA1 ($k = 1$) model to the data using the covariance parameterization.



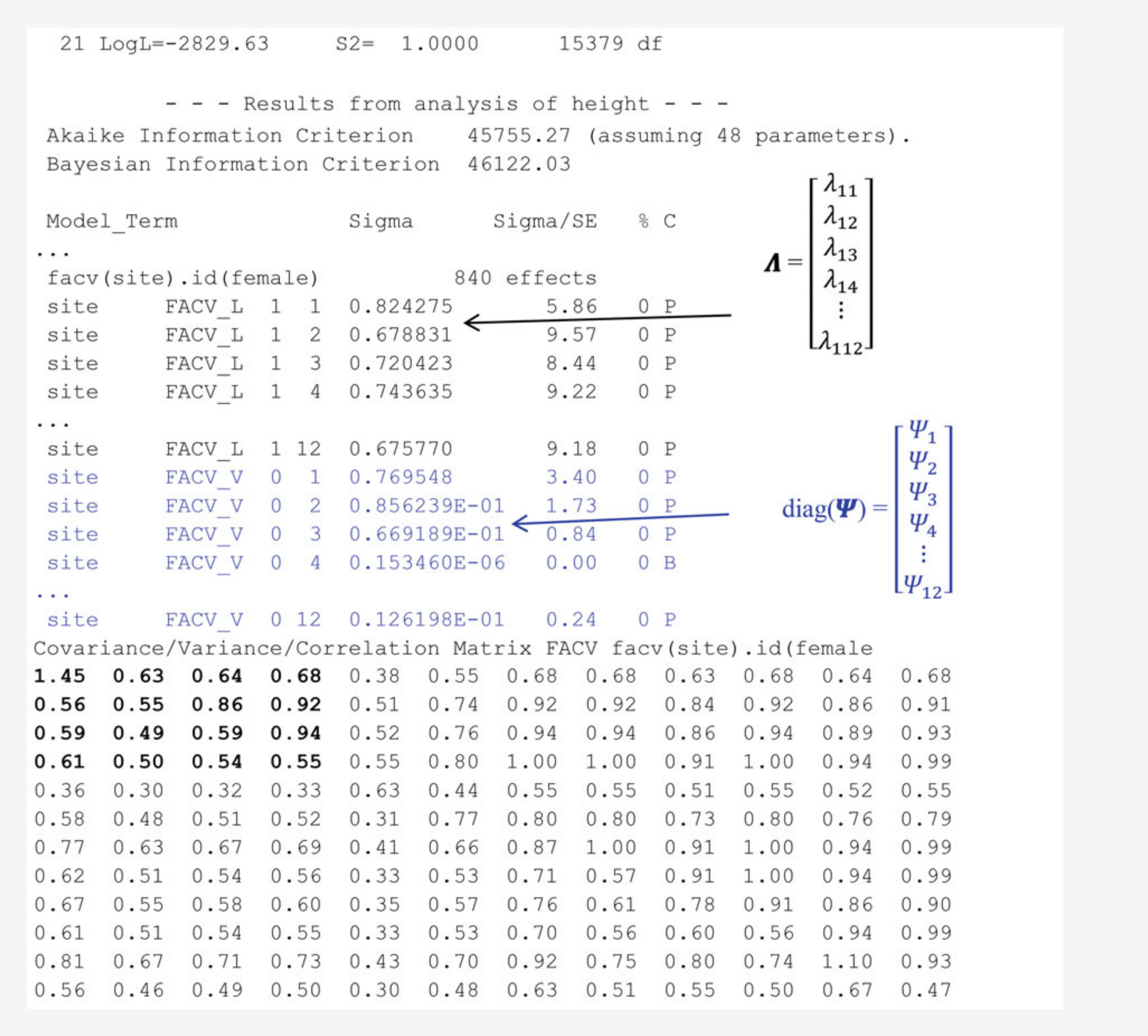
```
!PART 9
$B ~ mu site !r  ,
facv(site).id(female) ,
 idh(site).block
     residual sat(site).id(units)
```

$$\begin{bmatrix} \lambda_{11}^2+\Psi_1 & \lambda_{11}\lambda_{12} & \lambda_{11}\lambda_{13} & \lambda_{11}\lambda_{14} \\ \lambda_{11}\lambda_{12} & \lambda_{12}^2+\Psi_2 & \lambda_{12}\lambda_{13} & \lambda_{12}\lambda_{14} \\ \lambda_{11}\lambda_{13} & \lambda_{12}\lambda_{13} & \lambda_{13}^2+\Psi_3 & \lambda_{13}\lambda_{14} \\ \lambda_{11}\lambda_{14} & \lambda_{12}\lambda_{14} & \lambda_{13}\lambda_{14} & \lambda_{14}^2+\Psi_4 \end{bmatrix} \otimes \mathbf{I}_m$$

- The variance-covariance structure for the compound term `site.female` is the direct product of an FA1 matrix for site effects and an identity matrix for female effects. If pedigree information were available on the females, we could use `nrm (female)` to account for genetic relationships among females.

A subset of the *.asr* output file is given here:

OUTPUT 9: FA1 covariance model



```
  21 LogL=-2829.63     S2=  1.0000      15379 df

         - - - Results from analysis of height - - -
 Akaike Information Criterion    45755.27 (assuming 48 parameters).
 Bayesian Information Criterion  46122.03

 Model_Term                Sigma      Sigma/SE   % C
...
 facv(site).id(female)            840 effects
 site      FACV_L  1  1  0.824275         5.86   0 P
 site      FACV_L  1  2  0.678831         9.57   0 P
 site      FACV_L  1  3  0.720423         8.44   0 P
 site      FACV_L  1  4  0.743635         9.22   0 P
...
 site      FACV_L  1 12  0.675770         9.18   0 P
 site      FACV_V  0  1  0.769548         3.40   0 P
 site      FACV_V  0  2  0.856239E-01     1.73   0 P
 site      FACV_V  0  3  0.669189E-01     0.84   0 P
 site      FACV_V  0  4  0.153460E-06     0.00   0 B
...
 site      FACV_V  0 12  0.126198E-01     0.24   0 P
Covariance/Variance/Correlation Matrix FACV facv(site).id(female
1.45  0.63  0.64  0.68  0.38  0.55  0.68  0.68  0.63  0.68  0.64  0.68
0.56  0.55  0.86  0.92  0.51  0.74  0.92  0.92  0.84  0.92  0.86  0.91
0.59  0.49  0.59  0.94  0.52  0.76  0.94  0.94  0.86  0.94  0.89  0.93
0.61  0.50  0.54  0.55  0.55  0.80  1.00  1.00  0.91  1.00  0.94  0.99
0.36  0.30  0.32  0.33  0.63  0.44  0.55  0.55  0.51  0.55  0.52  0.55
0.58  0.48  0.51  0.52  0.31  0.77  0.80  0.80  0.73  0.80  0.76  0.79
0.77  0.63  0.67  0.69  0.41  0.66  0.87  1.00  0.91  1.00  0.94  0.99
0.62  0.51  0.54  0.56  0.33  0.53  0.71  0.57  0.91  1.00  0.94  0.99
0.67  0.55  0.58  0.60  0.35  0.57  0.76  0.61  0.78  0.91  0.86  0.90
0.61  0.51  0.54  0.55  0.33  0.53  0.70  0.56  0.60  0.56  0.94  0.99
0.81  0.67  0.71  0.73  0.43  0.70  0.92  0.75  0.80  0.74  1.10  0.93
0.56  0.46  0.49  0.50  0.30  0.48  0.63  0.51  0.55  0.50  0.67  0.47
```

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{14} \\ \vdots \\ \lambda_{112} \end{bmatrix}$$

$$\mathrm{diag}(\boldsymbol{\Psi}) = \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \\ \vdots \\ \Psi_{12} \end{bmatrix}$$

- The log likelihood of the FA1 model is −2829.63, which is similar to the `corgh()` model (model 8 log likelihood = −2804.06), although the FA1 model requires only 24 parameters for the genotype within environment covariance matrix compared to 78 for the `us()`/`corgh()` models. By capturing the variance/covariance structure well with many fewer parameters, the FA1 model has much better (lower) Akaike and Bayesian Information Criteria than the `corgh()` model.
- In the *.asr* output file, site loadings on the correlation scale are labeled 'FACV_L'. Values with label 'FACV_V' are the site-specific genetic variances (the diagonal elements of Ψ).
- The within-site genetic variances and between site covariances and correlation estimates are given in the 'covariance/variance/correlation matrix' at the bottom of the output. In the example output above, we highlighted in bold the estimates for the first four environments.
- The diagonal elements of the FACV covariance matrix are obtained as the squared loadings plus the site-specific variances. For example, for site 1, the variance (element [1,1] in the covariance/variance/correlation matrix) is:

$$\sigma^2_{g(e)1} = \lambda^2_{11} + \Psi_1 \quad = (0.824275)^2 + 0.769548 = 1.45$$

- Notice that a relatively large additional site-specific variance must be added to the squared loading for site 1 to obtain a good estimate of the within-site genetic variance. In contrast, for site 4, its within-site variance is estimated accurately by the square of its loading, so its site-specific variance is close to zero.
- The estimated genetic covariance between a family's performance at sites 1 and 2 (element [2,1] in the covariance/variance/correlation matrix) is simply the product of their loadings:

$$\sigma_{g12} = \lambda_{11}\lambda_{12} = (0.824275)(0.678831) = 0.56$$

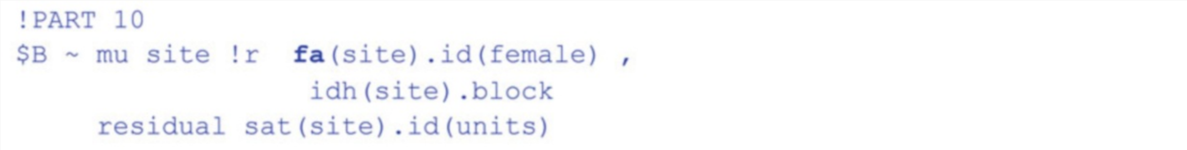
- The estimated genetic correlation between a family's performance at sites 1 and 2 (element [2,1] in the covariance/variance/correlation matrix) is the covariance divided by the square root of the product of the within-site genetic variances:

$$r_{g12} = \frac{\lambda_{11}\lambda_{12}}{\sqrt{\left(\lambda^2_{11} + \Psi_1\right)\left(\lambda^2_{12} + \Psi_2\right)}} = \frac{0.56}{\sqrt{(1.45)(0.55)}} = 0.62$$

- In this example, female effects represent half-sib family means, so the genetic variance and covariance estimates are a quarter of the additive genetic variances/covariances. The correlation estimates are additive genetic correlations.

Model 10: FA1 Correlation Structure

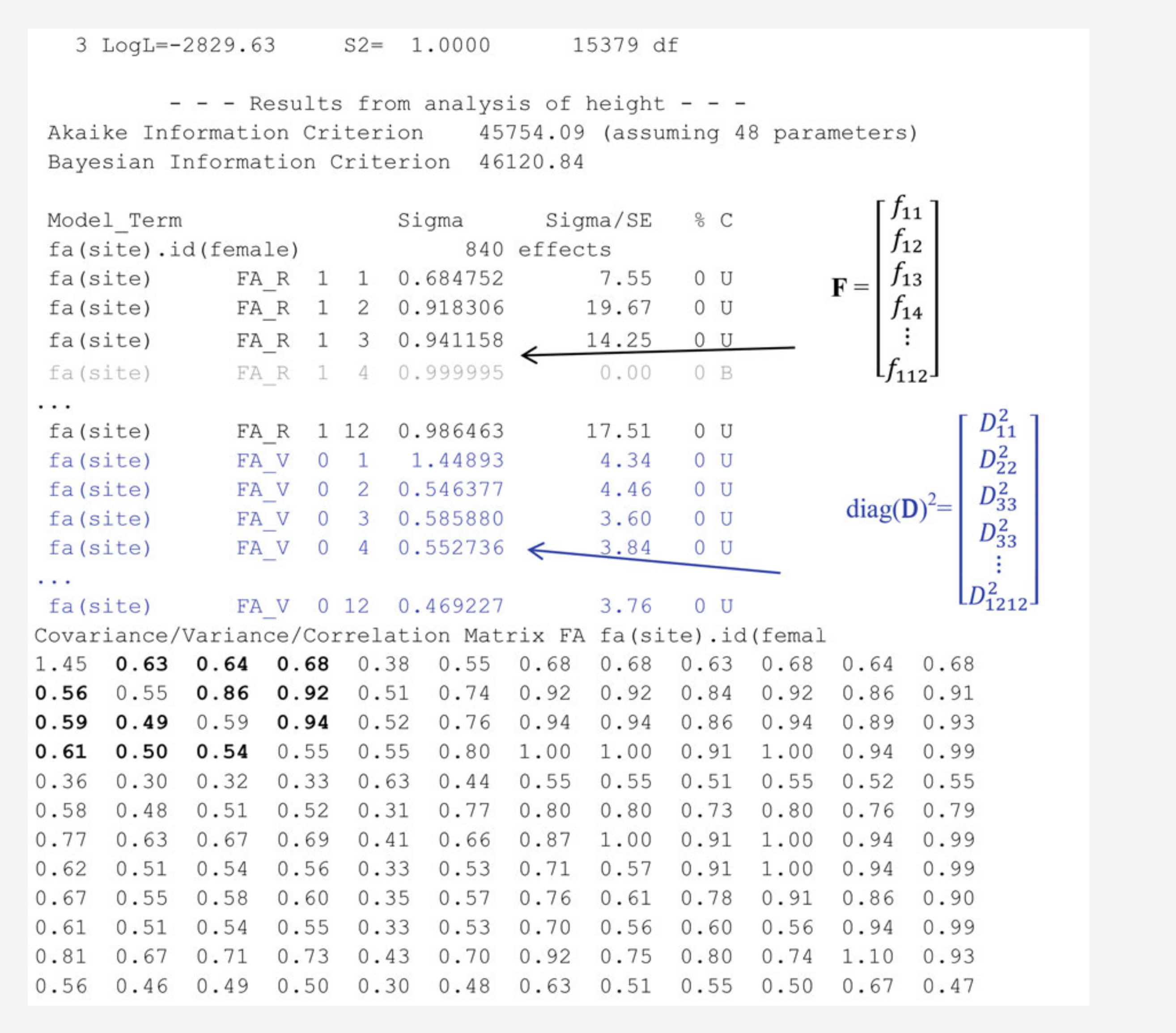
In PART 10 we fit the FA1 model using the correlation parameterization.



```
!PART 10
$B ~ mu site !r  fa(site).id(female) ,
                  idh(site).block
     residual sat(site).id(units)
```

A subset of the *.asr* output file is given here:

OUTPUT 10: FA1 correlation model



```
  3 LogL=-2829.63     S2=  1.0000      15379 df

         - - - Results from analysis of height - - -
 Akaike Information Criterion    45754.09 (assuming 48 parameters)
 Bayesian Information Criterion  46120.84

 Model_Term                 Sigma      Sigma/SE   % C
 fa(site).id(female)            840 effects
 fa(site)        FA_R  1  1  0.684752       7.55   0 U
 fa(site)        FA_R  1  2  0.918306      19.67   0 U
 fa(site)        FA_R  1  3  0.941158      14.25   0 U
 fa(site)        FA_R  1  4  0.999995       0.00   0 B
...
 fa(site)        FA_R  1 12  0.986463      17.51   0 U
 fa(site)        FA_V  0  1   1.44893       4.34   0 U
 fa(site)        FA_V  0  2  0.546377       4.46   0 U
 fa(site)        FA_V  0  3  0.585880       3.60   0 U
 fa(site)        FA_V  0  4  0.552736       3.84   0 U
...
 fa(site)        FA_V  0 12  0.469227       3.76   0 U
Covariance/Variance/Correlation Matrix FA fa(site).id(femal
1.45  0.63  0.64  0.68  0.38  0.55  0.68  0.68  0.63  0.68  0.64  0.68
0.56  0.55  0.86  0.92  0.51  0.74  0.92  0.92  0.84  0.92  0.86  0.91
0.59  0.49  0.59  0.94  0.52  0.76  0.94  0.94  0.86  0.94  0.89  0.93
0.61  0.50  0.54  0.55  0.55  0.80  1.00  1.00  0.91  1.00  0.94  0.99
0.36  0.30  0.32  0.33  0.63  0.44  0.55  0.55  0.51  0.55  0.52  0.55
0.58  0.48  0.51  0.52  0.31  0.77  0.80  0.80  0.73  0.80  0.76  0.79
0.77  0.63  0.67  0.69  0.41  0.66  0.87  1.00  0.91  1.00  0.94  0.99
0.62  0.51  0.54  0.56  0.33  0.53  0.71  0.57  0.91  1.00  0.94  0.99
0.67  0.55  0.58  0.60  0.35  0.57  0.76  0.61  0.78  0.91  0.86  0.90
0.61  0.51  0.54  0.55  0.33  0.53  0.70  0.56  0.60  0.56  0.94  0.99
0.81  0.67  0.71  0.73  0.43  0.70  0.92  0.75  0.80  0.74  1.10  0.93
0.56  0.46  0.49  0.50  0.30  0.48  0.63  0.51  0.55  0.50  0.67  0.47
```

$$\mathbf{F}=\begin{bmatrix} f_{11}\\ f_{12}\\ f_{13}\\ f_{14}\\ \vdots\\ f_{112}\end{bmatrix}$$

$$\text{diag}(\mathbf{D})^2=\begin{bmatrix} D_{11}^2\\ D_{22}^2\\ D_{33}^2\\ D_{33}^2\\ \vdots\\ D_{1212}^2\end{bmatrix}$$

- In the *.asr* output file, residual variances and genetic correlations for pairs of sites and genetic variances are reported (under the column heading ‘sigma’). Site loadings on the correlation scale are labeled ‘FA_R’ in the output. Values with label ‘FA_V’ are genetic variances within each site (which are the sum of the squared site loadings and the site-specific variance). Several site loadings on the correlation scale are very close to one (FA_R = 0.9995) and are constrained at the boundary flagged by ‘B’.
- The genetic variance and correlation estimates are also given in the covariance/variance/correlation matrix at the bottom of the output. Notice that the model likelihood and the covariance/variance/correlation estimates are identical for models 9 and 10. The only difference is in how the parameter estimates are reported.
- The loadings on the correlation scale are equal to the covariance model loadings divided by the square root of the within-site genetic variance. For example, for site 1, the correlation loading is equated to the covariance model parameters as:

$$f_{11}=\frac{\lambda_{11}}{\sqrt{\lambda_{11}^2+\Psi_1}}=\frac{0.824275}{\sqrt{1.45}}=0.68$$

- The estimates labelled as ‘FA_V’ are the squared diagonal elements of the **D** matrix, equal to the within-site variances estimated from the covariance parameterization. For example, for site 1: $D_{11}^2=\lambda_{11}^2+\Psi_1=1.45$

The between-site genetic correlations are obtained directly as products of the correlation loadings (the 'FA_R' values in the output), which are the elements of the **F** vector. As an example, consider the loadings for only the first four environments:

$$\mathbf{F} = \begin{bmatrix} 0.6847 \\ 0.9183 \\ 0.9412 \\ 0.9999 \end{bmatrix}$$

We can construct something close to the correlation matrix from the product $\mathbf{FF}^{\mathrm{T}}$.

$$\boldsymbol{FF}^T = \begin{bmatrix} 0.47 & 0.63 & 0.64 & 0.68 \\ 0.63 & 0.84 & 0.86 & 0.92 \\ 0.64 & 0.86 & 0.89 & 0.94 \\ 0.68 & 0.92 & 0.94 & 0.99 \end{bmatrix}$$

The off-diagonal elements of the product are the correlations between pairs of environments, e.g.

$r_{12} = 0.6847 * 0.9183 = 0.63.$

However, the diagonal elements are not equal to one, so $\mathbf{FF}^{\mathrm{T}}$ is not a proper correlation matrix. For example, the element (1,1) of $\mathbf{FF}^{\boldsymbol{T}}$ is $(0.6847)^2 = 0.47$. Therefore we construct a matrix $\mathbf{E} = \mathrm{diag}(\mathbf{1} - \mathbf{F}^2)$ and add it to $\mathbf{FF}^{\mathrm{T}}$ to make the correlation matrix **C**, which now has diagonal elements equal to exactly one:

$$\mathbf{E} = \begin{bmatrix} 1-(0.6847)^2 & 0 & 0 & 0 \\ 0 & 1-(0.9183)^2 & 0 & 0 \\ 0 & 0 & 1-(0.9412)^2 & 0 \\ 0 & 0 & 0 & 1-(0.9999)^2 \end{bmatrix}$$

$$\mathbf{C} = \boldsymbol{FF}^T + \boldsymbol{E} = \begin{bmatrix} 1 & 0.63 & 0.64 & 0.68 \\ 0.63 & 1 & 0.86 & 0.92 \\ 0.64 & 0.86 & 1 & 0.94 \\ 0.68 & 0.92 & 0.94 & 1 \end{bmatrix}$$

The **D** matrix has square roots of the genetic variances within each site on the diagonal:

$$\mathbf{D} = \begin{bmatrix} \sqrt{1.45} & 0 & 0 & 0 \\ 0 & \sqrt{0.55} & 0 & 0 \\ 0 & 0 & \sqrt{0.59} & 0 \\ 0 & 0 & 0 & \sqrt{0.55} \end{bmatrix}$$

The correlation matrix for family within site effects is obtained as:

$$\mathbf{G} = \mathbf{DCD}^{\boldsymbol{T}} \otimes \boldsymbol{I\sigma}_{\boldsymbol{F}}^2 = \begin{bmatrix} \mathbf{1.45} & 0.56 & 0.60 & 0.61 \\ 0.56 & \mathbf{0.55} & 0.49 & 0.51 \\ 0.60 & 0.49 & \mathbf{0.59} & 0.54 \\ 0.61 & 0.51 & 0.54 & \mathbf{0.55} \end{bmatrix} \otimes \boldsymbol{I\sigma}_{\boldsymbol{F}}^2$$

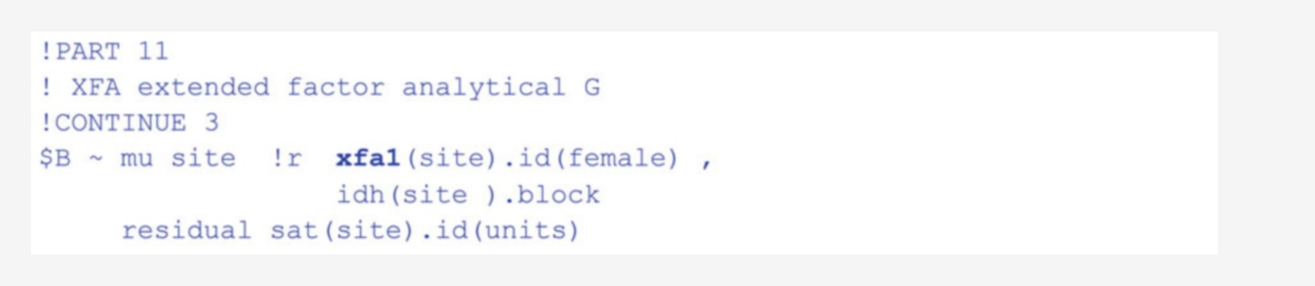
Now, consider how this model can be reformulated in terms of a covariance matrix. The loadings on the covariance scale (**Λ**) are equal to the product **DF** from the correlation parameterization:

$$\mathbf{\Lambda} = \mathbf{DF} = \begin{bmatrix} \sqrt{1.45} & 0 & 0 & 0 \\ 0 & \sqrt{0.55} & 0 & 0 \\ 0 & 0 & \sqrt{0.59} & 0 \\ 0 & 0 & 0 & \sqrt{0.55} \end{bmatrix} \begin{bmatrix} 0.6847 \\ 0.9183 \\ 0.9412 \\ 0.9999 \end{bmatrix} = \begin{bmatrix} 0.8243 \\ 0.6788 \\ 0.7204 \\ 0.7436 \end{bmatrix}$$

This is the set of loadings we obtained with the covariance forms of the FA1 model (model 9).

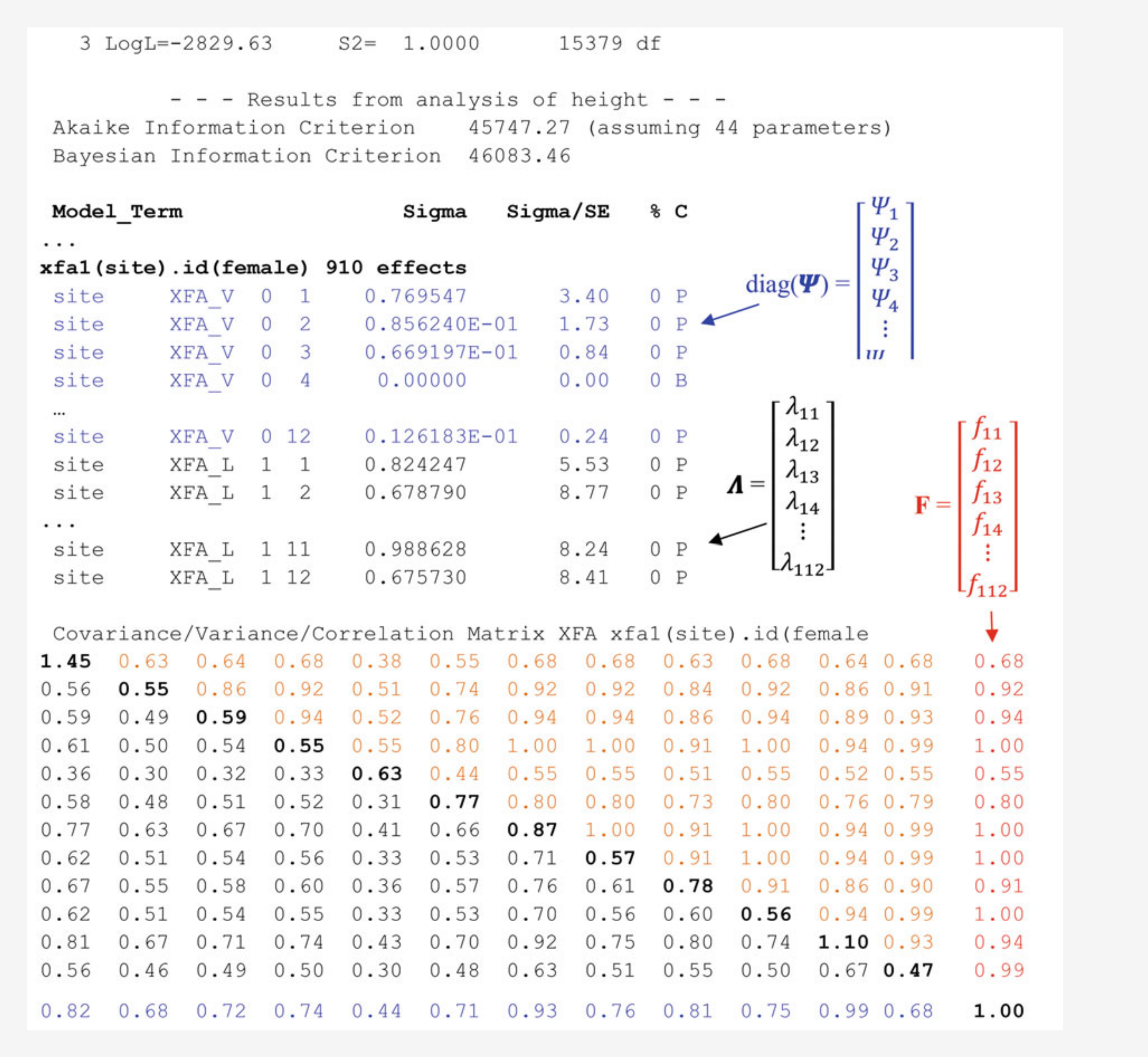
Model 11: XFA*1* Structure

The XFA1 model is a third model equivalent to FACV1 and FA1, but has a different parameterization that improves computational efficiency.



```
!PART 11
! XFA extended factor analytical G
!CONTINUE 3
$B ~ mu site  !r  xfa1(site).id(female) ,
                  idh(site ).block
     residual sat(site).id(units)
```

OUTPUT 10: A subset of the output from XFA1 model



```
   3 LogL=-2829.63     S2=  1.0000      15379 df

          - - - Results from analysis of height - - -
 Akaike Information Criterion    45747.27 (assuming 44 parameters)
 Bayesian Information Criterion  46083.46

 Model_Term                  Sigma   Sigma/SE   % C
...
xfa1(site).id(female) 910 effects
 site    XFA_V  0  1   0.769547       3.40    0 P
 site    XFA_V  0  2   0.856240E-01   1.73    0 P
 site    XFA_V  0  3   0.669197E-01   0.84    0 P
 site    XFA_V  0  4    0.00000       0.00    0 B
 …
 site    XFA_V  0 12   0.126183E-01   0.24    0 P
 site    XFA_L  1  1   0.824247       5.53    0 P
 site    XFA_L  1  2   0.678790       8.77    0 P
...
 site    XFA_L  1 11   0.988628       8.24    0 P
 site    XFA_L  1 12   0.675730       8.41    0 P

 Covariance/Variance/Correlation Matrix XFA xfa1(site).id(female
1.45  0.63  0.64  0.68  0.38  0.55  0.68  0.68  0.63  0.68  0.64 0.68   0.68
0.56  0.55  0.86  0.92  0.51  0.74  0.92  0.92  0.84  0.92  0.86 0.91   0.92
0.59  0.49  0.59  0.94  0.52  0.76  0.94  0.94  0.86  0.94  0.89 0.93   0.94
0.61  0.50  0.54  0.55  0.55  0.80  1.00  1.00  0.91  1.00  0.94 0.99   1.00
0.36  0.30  0.32  0.33  0.63  0.44  0.55  0.55  0.51  0.55  0.52 0.55   0.55
0.58  0.48  0.51  0.52  0.31  0.77  0.80  0.80  0.73  0.80  0.76 0.79   0.80
0.77  0.63  0.67  0.70  0.41  0.66  0.87  1.00  0.91  1.00  0.94 0.99   1.00
0.62  0.51  0.54  0.56  0.33  0.53  0.71  0.57  0.91  1.00  0.94 0.99   1.00
0.67  0.55  0.58  0.60  0.36  0.57  0.76  0.61  0.78  0.91  0.86 0.90   0.91
0.62  0.51  0.54  0.55  0.33  0.53  0.70  0.56  0.60  0.56  0.94 0.99   1.00
0.81  0.67  0.71  0.74  0.43  0.70  0.92  0.75  0.80  0.74  1.10 0.93   0.94
0.56  0.46  0.49  0.50  0.30  0.48  0.63  0.51  0.55  0.50  0.67 0.47   0.99

0.82  0.68  0.72  0.74  0.44  0.71  0.93  0.76  0.81  0.75  0.99 0.68   1.00
```

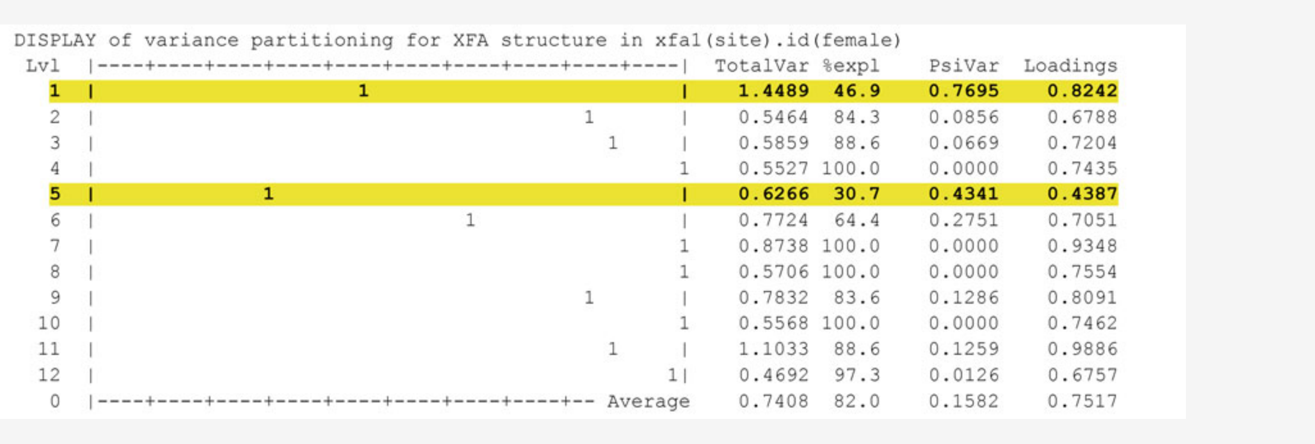
$$\mathrm{diag}(\boldsymbol{\Psi}) = \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \\ \vdots \\ \Psi \end{bmatrix}$$

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{14} \\ \vdots \\ \lambda_{112} \end{bmatrix}$$

$$\mathbf{F} = \begin{bmatrix} f_{11} \\ f_{12} \\ f_{13} \\ f_{14} \\ \vdots \\ f_{112} \end{bmatrix}$$

- The residual LogL is the same as it was for models 9 and 10. The AIC/BIC values of model 11 are different because ASReml does not count any site-specific variances in Ψ that are fixed at zero as parameters in the XFA model, whereas these parameters are set to very small values close to zero in the FACV and FA models. This artificially makes the XFA1 appear to be a better fitting model, but effectively the models are all the same.
- The parameter estimates in the output for the XFA model are identical to the FACV model, except they appear in different order.
- Parameter estimates labeled ‘XFA_V’ are the site-specific variances (the diagonal elements of Ψ), four of which are fixed at 0 in this example.
- The values labeled ‘XFA_L’ are site loadings on the covariance scale (**Λ**).
- The covariance/variance/correlation matrix for sites is given at the bottom of the output. This matrix is identical to the matrices estimated by models 9 and 10, except that one extra row and one extra column are added to the matrix.
- The extra column added to the right side of the matrix contains the factor loadings on the correlation scale (equal to the **F** vector in the FA model).
- The additional row at the bottom of the matrix has the factor loadings on the covariance scale (**Λ**).

To aid with model diagnosis and selection, a plot of the proportion of within-site variances estimated by the factor part of the model appears in the *.res* file (“Code 8-1_MET11_height.res”). The column labeled “%expl” corresponds to the within-site genetic variances described in Eqs. 8.12 and 8.13:

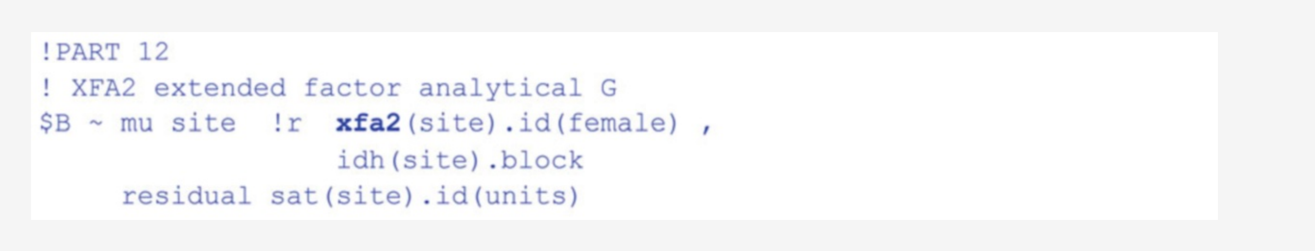


```
DISPLAY of variance partitioning for XFA structure in xfa1(site).id(female)
 Lvl  |---+---+---+---+---+---+---+---+---+---|  TotalVar %expl    PsiVar  Loadings
  1  |                 1                     |    1.4489  46.9    0.7695    0.8242
  2  |                                 1     |    0.5464  84.3    0.0856    0.6788
  3  |                                  1    |    0.5859  88.6    0.0669    0.7204
  4  |                                       1    0.5527 100.0    0.0000    0.7435
  5  |           1                           |    0.6266  30.7    0.4341    0.4387
  6  |                         1             |    0.7724  64.4    0.2751    0.7051
  7  |                                       1    0.8738 100.0    0.0000    0.9348
  8  |                                       1    0.5706 100.0    0.0000    0.7554
  9  |                                 1     |    0.7832  83.6    0.1286    0.8091
 10  |                                       1    0.5568 100.0    0.0000    0.7462
 11  |                                  1    |    1.1033  88.6    0.1259    0.9886
 12  |                                      1|    0.4692  97.3    0.0126    0.6757
  0  |---+---+---+---+---+---+---+---+-- Average  0.7408  82.0    0.1582    0.7517
```

In above output less than half of the variation among females within sites 1 and 5 (highlighted in the output) is explained by the factor part of the model, so fairly large site-specific variances (“PsiVar”) are needed to explain the observed variation at those environments.

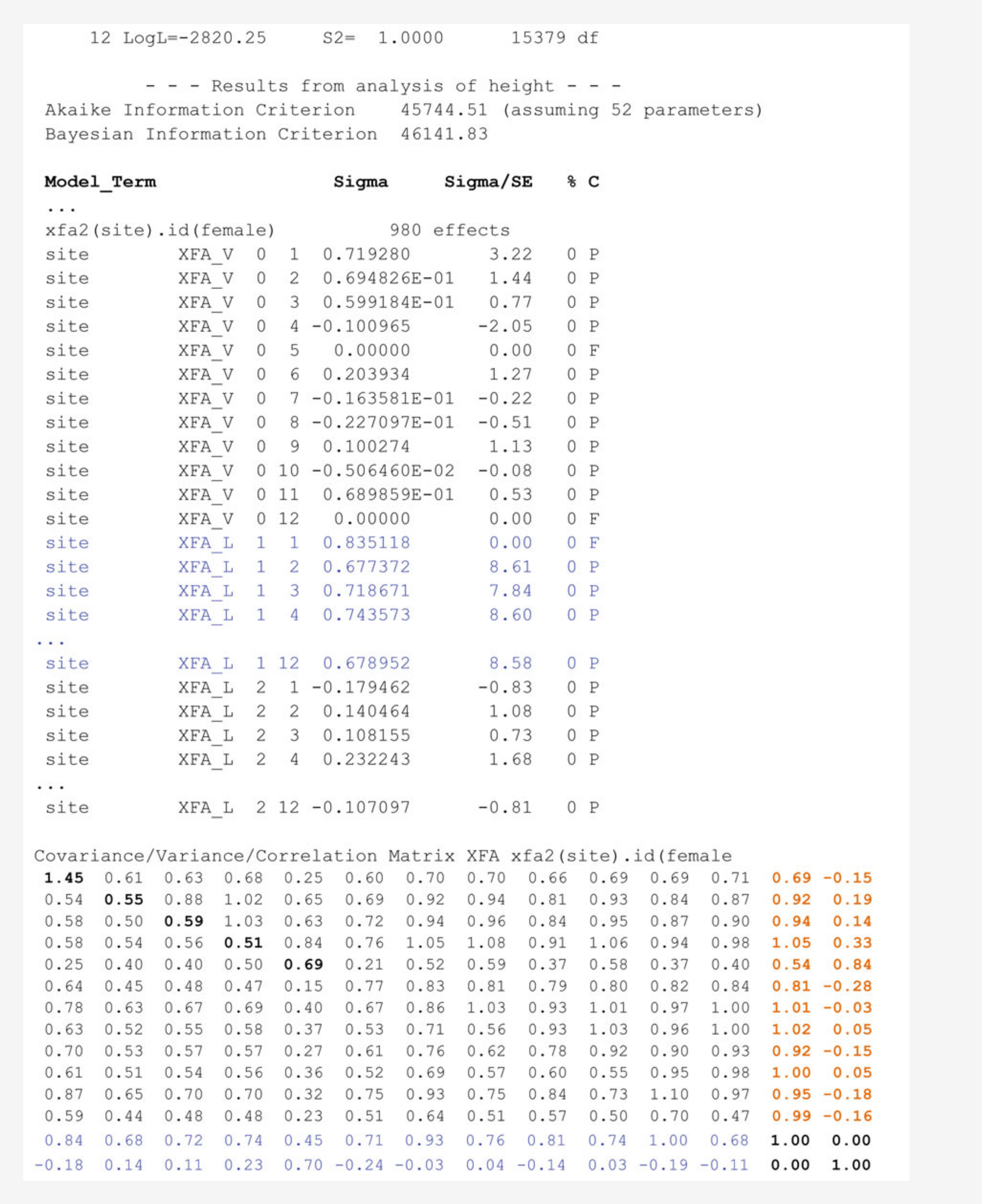
Model 12: XFA2 Structure

In PART 12 the *XFA with 2 factors* is fitted. The XFA2 model assumes that two factors explain the correlation structure between pairs of sites:



```
!PART 12
! XFA2 extended factor analytical G
$B ~ mu site  !r  xfa2(site).id(female) ,
                  idh(site).block
     residual sat(site).id(units)
```

The output of the XFA2 model follows (result may differ slightly due to different starting values).



```
     12 LogL=-2820.25     S2=  1.0000      15379 df

          - - - Results from analysis of height - - -
 Akaike Information Criterion    45744.51 (assuming 52 parameters)
 Bayesian Information Criterion  46141.83

 Model_Term                  Sigma     Sigma/SE   % C
 ...
 xfa2(site).id(female)            980 effects
 site         XFA_V  0  1  0.719280        3.22   0 P
 site         XFA_V  0  2  0.694826E-01    1.44   0 P
 site         XFA_V  0  3  0.599184E-01    0.77   0 P
 site         XFA_V  0  4 -0.100965       -2.05   0 P
 site         XFA_V  0  5   0.00000        0.00   0 F
 site         XFA_V  0  6  0.203934        1.27   0 P
 site         XFA_V  0  7 -0.163581E-01   -0.22   0 P
 site         XFA_V  0  8 -0.227097E-01   -0.51   0 P
 site         XFA_V  0  9  0.100274        1.13   0 P
 site         XFA_V  0 10 -0.506460E-02   -0.08   0 P
 site         XFA_V  0 11  0.689859E-01    0.53   0 P
 site         XFA_V  0 12   0.00000        0.00   0 F
 site         XFA_L  1  1  0.835118        0.00   0 F
 site         XFA_L  1  2  0.677372        8.61   0 P
 site         XFA_L  1  3  0.718671        7.84   0 P
 site         XFA_L  1  4  0.743573        8.60   0 P
...
 site         XFA_L  1 12  0.678952        8.58   0 P
 site         XFA_L  2  1 -0.179462       -0.83   0 P
 site         XFA_L  2  2  0.140464        1.08   0 P
 site         XFA_L  2  3  0.108155        0.73   0 P
 site         XFA_L  2  4  0.232243        1.68   0 P
...
 site         XFA_L  2 12 -0.107097       -0.81   0 P

Covariance/Variance/Correlation Matrix XFA xfa2(site).id(female
 1.45  0.61  0.63  0.68  0.25  0.60  0.70  0.70  0.66  0.69  0.69  0.71  0.69 -0.15
 0.54  0.55  0.88  1.02  0.65  0.69  0.92  0.94  0.81  0.93  0.84  0.87  0.92  0.19
 0.58  0.50  0.59  1.03  0.63  0.72  0.94  0.96  0.84  0.95  0.87  0.90  0.94  0.14
 0.58  0.54  0.56  0.51  0.84  0.76  1.05  1.08  0.91  1.06  0.94  0.98  1.05  0.33
 0.25  0.40  0.40  0.50  0.69  0.21  0.52  0.59  0.37  0.58  0.37  0.40  0.54  0.84
 0.64  0.45  0.48  0.47  0.15  0.77  0.83  0.81  0.79  0.80  0.82  0.84  0.81 -0.28
 0.78  0.63  0.67  0.69  0.40  0.67  0.86  1.03  0.93  1.01  0.97  1.00  1.01 -0.03
 0.63  0.52  0.55  0.58  0.37  0.53  0.71  0.56  0.93  1.03  0.96  1.00  1.02  0.05
 0.70  0.53  0.57  0.57  0.27  0.61  0.76  0.62  0.78  0.92  0.90  0.93  0.92 -0.15
 0.61  0.51  0.54  0.56  0.36  0.52  0.69  0.57  0.60  0.55  0.95  0.98  1.00  0.05
 0.87  0.65  0.70  0.70  0.32  0.75  0.93  0.75  0.84  0.73  1.10  0.97  0.95 -0.18
 0.59  0.44  0.48  0.48  0.23  0.51  0.64  0.51  0.57  0.50  0.70  0.47  0.99 -0.16
 0.84  0.68  0.72  0.74  0.45  0.71  0.93  0.76  0.81  0.74  1.00  0.68  1.00  0.00
-0.18  0.14  0.11  0.23  0.70 -0.24 -0.03  0.04 -0.14  0.03 -0.19 -0.11  0.00  1.00
```

- The XFA2 model has a better log likelihood than the XFA1 model (−2820.25 vs. −2829.63) but it uses 12 additional parameters to capture additional variation. Depending on the penalty used for adding parameters to the model, the XFA2 model could be considered better or worse than the XFA1 model. The XFA2 model has better Akaike Information Criterion than the XFA1 model (45744.51 for XFA2 vs. 45747.27 for XFA1) but worse Bayesian Information Criterion

(46141.83 for XFA2 vs. 46083.46 for XFA1). Therefore, choice of XFA1 vs XFA2 model in this case is not clear cut and is up to the judgement of the researcher.

- Parameter estimates labeled 'XFA_V' are the site-specific variances and 'XFA_L' are loadings. For the XFA2 model, the loadings are indexed by the factor number (1 or 2) and the site number (1 through 12):
  – `XFA_L 1 1` refers to the loading on the first factor for the first site,
  – `XFA_L 1 2` refers to the loading on the first factor for the second site,
  – `XFA_L 2 1` refers to the loading on the second factor for the first site and so forth.
- For the XFA2 model, $\mathbf{\Lambda}_g$ has s rows for sites and two columns for two factors. The loadings for the first four environments on the two factors are:

$$\mathbf{\Lambda} = \begin{bmatrix} 0.84 & -0.18 \\ 0.68 & -0.14 \\ 0.72 & 0.11 \\ 0.74 & 0.23 \end{bmatrix}$$

- Notice that the loadings on the first factor are different than the loadings in XFA1 model. For the second factor, some sites had negative loadings. The factor loadings are not unique solutions, and other solutions can be produced.
- The last two columns in the XFA output (orange color vectors) are site loadings on the correlation scale. Notice that correlations can go out of theoretical bounds (>1) in the XFA2 model.
- Also notice that some of the site-specific variances (for example, site 4) are negative. The genetic variance predicted at site 4 based on the two factors is the sum of the squared loadings for site 4: $\sum_{r=1}^{2} \lambda_{i4}^2 = (0.74)^2 + (0.23)^2 = 0.60$. However, this is an overestimate of the genetic variance within site 4. So, a negative site-specific variance needs to be added to the sum of squared loadings to get a better estimate of the within-site variance: $Var\left(G(E)_4\right) = \sum_{r=1}^{2} \lambda_{i4}^2 + \Psi_4 = 0.60 - 0.10 = 0.50$. This is within rounding error of element [4,4] of the covariance/variance/correlation matrix in the output above.

Model 13: XFA*3* Structure

In PART 13 of the example code, the *XFA with 3 factors* model is fitted. We do not show the output from this model, as its AIC and BIC values are worse than the XFA2 model. A summary of the models fit to pine polymix data is given in Table 8.1:

Model LogL values decrease as model complexity (the number of parameters) increases. AIC value follows the same trend until the FA2 model, after which the penalty for additional parameters outweighs the improvement in likelihood. The BIC penalizes additional parameters more stringently, such that the simple CORUV model (equivalent to the classical factorial model) has the best BIC value. In such situations, model choice is not clear cut, but we note that the FA1/XFA1 model provides a good compromise between model fit and number of parameters, such that it has second best AIC and third best BIC.

MET Models with ASReml-R

For interested readers, an R markdown file (*Code 8-2_pine_met.Rmd*) and its knitted output (*Code 8-2_pine_met.html*) are provided to show the sequence of analyses using ASReml-R. In ASReml-R, the US and FA3 models did not converge despite using initial values and `update.asreml()` function of ASReml-R. Another detail that ASReml-R users should be aware of is that for FA models with $k > 1$, the factor loading solutions are not unique and ASReml-R produces different

Table 8.1 Model fit statistics (log likelihood, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), standard error of differences of family mean predictions, and number of parameters for the site.female term) for pine polymix data

Model G Structure	LogL	AIC	BIC	SED	G(E) Parameters
CORUV (Model 5)	−2865.30	45782.59	**45981.25**	0.2418	2
CORUH (Model 6)	−2850.48	45774.97	46057.67	0.2425	13
CORGH (Model 8)	−2804.06	45812.12	46591.48	0.2395	78
XFA1 (Models 9/10/11)	−2829.04	45747.27	46083.46	0.2438	24
XFA2 (Model 12)	−2818.66	**45744.51**	46141.83	0.2413	56
XFA3 (Model 13)	−2814.42	45754.63	46236.20	0.2412	63

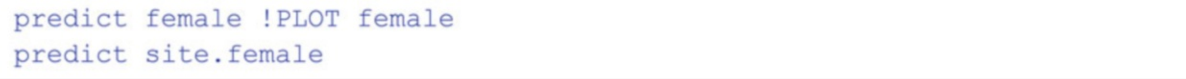
Models differ for site.female compound term only solutions than ASReml standalone (Cullis et al. 2010). We show in the code how to perform an orthogonal rotation of the ASReml-R loading solutions to match the ASReml standalone solutions.

Genetic Prediction with FA Models

From MET data, we can predict family values within sites or averaged across sites. Typically, the family means across sites are most useful for selection, but there are cases in which one site or group of sites may be distinct (with a low or negative correlation with other sites) and we may want to predict performance specifically in different groups of sites. Here, we demonstrate how to predict family values within and across sites with ASReml and how the predictions relate to the model effect estimates. We start with the simplest model (cross-classified and compound symmetry) and continue to the XFA1 model.

Model 3 – Cross-classified predictions

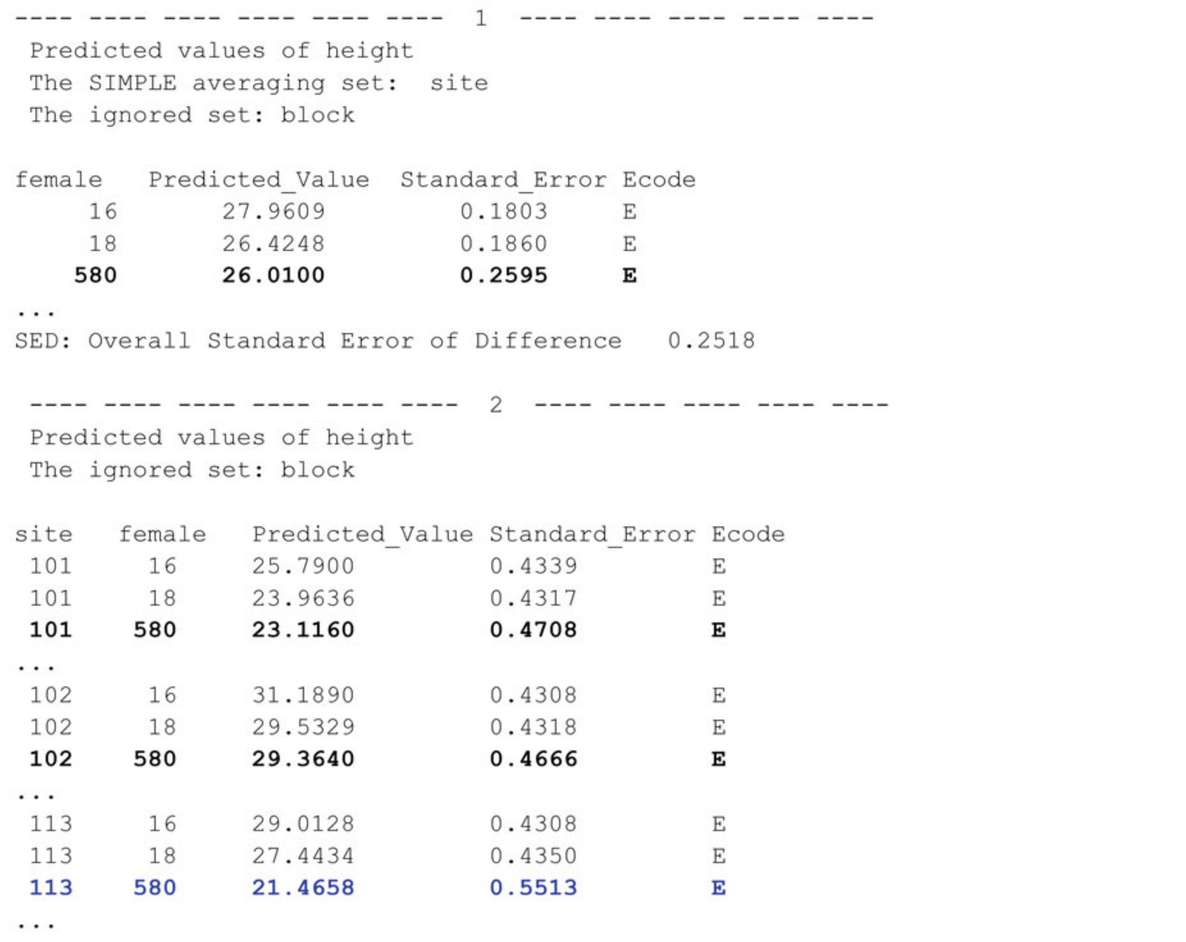
For the cross-classified model, we can obtain the across-site and within-site predictions using:



```
predict female !PLOT female
predict site.female
```

- The first predict statement is for prediction of female effect across all the sites. The second predict statement is site-specific predictions for females.
- The `!PLOT` qualifier produces a postscript graphic of predictions for females with one standard error.

The across-site predictions appear in the *.pvs* file above the site-specific predictions:



```
---- ---- ---- ---- ---- ----  1  ---- ---- ---- ---- ----
 Predicted values of height
 The SIMPLE averaging set:  site
 The ignored set: block

female   Predicted_Value  Standard_Error Ecode
     16       27.9609         0.1803     E
     18       26.4248         0.1860     E
    580       26.0100         0.2595     E
...
SED: Overall Standard Error of Difference   0.2518

 ---- ---- ---- ---- ---- ----  2  ---- ---- ---- ---- ----
 Predicted values of height
 The ignored set: block

site   female    Predicted_Value Standard_Error Ecode
 101     16      25.7900         0.4339         E
 101     18      23.9636         0.4317         E
 101    580      23.1160         0.4708         E
...
 102     16      31.1890         0.4308         E
 102     18      29.5329         0.4318         E
 102    580      29.3640         0.4666         E
...
 113     16      29.0128         0.4308         E
 113     18      27.4434         0.4350         E
 113    580      21.4658         0.5513         E
...
```

- There are two separate predictions in the *.pvs* file. The first output at the top is predicted breeding values of females across the sites.
- In the second part of the prediction we included the predictions for three families (16, 18, 580) at sites 101, 102, and 103. Female 580 did not have data at site 103 but its value is predicted. Notice that the standard error of this site-specific prediction for female 580 is 0.5513, higher than standard error of its predictions in sites 101 and 102. This is because in the absence of data for the particular site-family combination, the prediction is based on the main effects of the site and the family, with the predicted interaction effect exactly zero. Users may want to exclude such predictions from the *.pvs* file. This can be done by requesting the site-specific predictions with the qualifier **`!present site female`**.

```
predict site.female !present site female
```

Model 4 – CORUV predictions

Although we do not have family main effects in the CORUV model, we can nevertheless predict family values across sites as well as within sites using the same prediction statements as for the cross-classified model:

```
predict female
predict site.female !present site female
```

This produces predictions identical to the cross-classified model:

```
female   Predicted_Value   Standard_Error  Ecode
    16        27.9609           0.1803      E
    18        26.4248           0.1860      E
   580        26.0100           0.2595      E
...
SED: Overall Standard Error of Difference 0.2518

---- ---- ---- ---- ---- ---- 2 ---- ---- ---- ---- ---- ----
 Predicted values of height
 The ignored set: block
 Warning: 6 non-estimable [empty] cell(s) may be omitted from the table.

site  female  Predicted_Value Standard_Error Ecode
101     16    25.7900         0.4339         E
101     18    23.9636         0.4317         E
101    580    23.1160         0.4708         E

site  female  Predicted_Value Standard_Error Ecode
102     16    31.1890         0.4308         E
102     18    29.5329         0.4318         E
102    580    29.3640         0.4666         E
...
site  female  Predicted_Value Standard_Error Ecode
113     16    23.4069         0.4373         E
113     18    21.8089         0.4383         E
...
```

- Since we used "`!present site female`" there is no prediction for female 580 at site 113.

Model 11 – XFA1 predictions

The FA and XFA models partition the family within environment effects into a part due to the multiplicative interactions between factor loadings and family scores, and a second part due to site-specific genetic deviations for the family. This permits some flexibility in the across-sites and within-site family predictions, as the predictions can account for or ignore the site-specific genetic deviations. Recall that the predicted effect for genotype *j* at environment *i*, accounting for both the factor loadings and the site-specific effects for an FA*k* model is:

$$\widehat{u}_{gij} = \widehat{\lambda}_{1i}\widehat{f}_{1j} + \widehat{\lambda}_{2i}\widehat{f}_{2j} + \ldots + \widehat{\lambda}_{ki}\widehat{f}_{kj} + \widehat{\delta}_{ij} \tag{8.20}$$

This is a prediction with narrow inference: it is the family's effect in the specific environment *i* included in the experiment. The predicted value of the family also includes the intercept and site effect:

$$\widehat{Y}_{ij} = \mu + \widehat{S}_i + \widehat{u}_{gij} \tag{8.21}$$

We can also make a prediction of the family's site-specific value based only on the factors, ignoring the site-specific deviations:

$$\widehat{u}^*_{gij} = \widehat{\lambda}_{1i}\widehat{f}_{1j} + \widehat{\lambda}_{2i}\widehat{f}_{2j} + \ldots + \widehat{\lambda}_{ki}\widehat{f}_{kj} \tag{8.22}$$

$$\widehat{Y}^*_{ij} = \mu + \overline{\overline{\widehat{S}}}_{.} + \widehat{u}^*_{gij} \tag{8.23}$$

This type of prediction has a wider inference: it refers to the family's predicted effect in a future environment that is perfectly correlated with environment *i*.

Similarly, the predictions across sites can refer to the average performance across the set of environments actually included in the study:

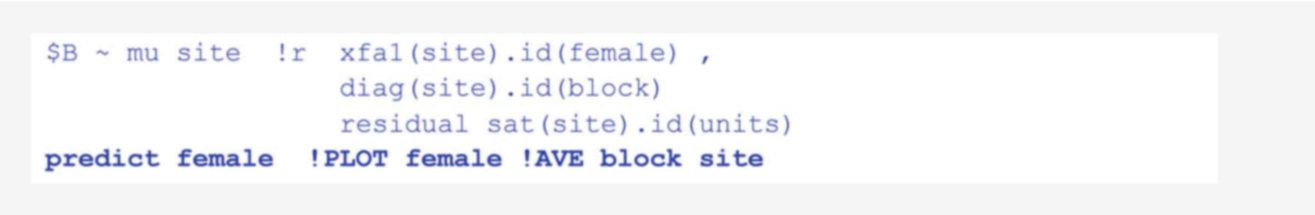
$$\widehat{\boldsymbol{Y}}_{.\boldsymbol{j}} = \boldsymbol{\mu} + \overline{\overline{\widehat{\boldsymbol{S}}}}_{.} + \overline{\widehat{\boldsymbol{u}}}_{\boldsymbol{g.j}} \tag{8.24}$$

This is equal to averaging the site-specific predictions including the site-specific genetic deviations. A prediction with wider inference would ignore the site-specific deviations and refer to performance at a hypothetical 'average' environment by predicting at the mean values of the factors:

$$\widehat{\boldsymbol{Y}}^*_{.\boldsymbol{j}} = \boldsymbol{\mu} + \overline{\overline{\widehat{\boldsymbol{S}}}}_{.} + \overline{\widehat{\boldsymbol{u}}}^*_{\boldsymbol{gij}} = \boldsymbol{\mu} + \overline{\overline{\widehat{\boldsymbol{S}}}}_{.} + \frac{1}{\boldsymbol{r}}\sum\nolimits_{\boldsymbol{k}=1}^{\boldsymbol{r}} \overline{\overline{\widehat{\boldsymbol{\lambda}}}}_{\boldsymbol{k}} \widehat{\boldsymbol{f}}_{\boldsymbol{kj}} \tag{8.25}$$

Here we demonstrate how to obtain these various predictions from ASReml, using the XFA1 model. In this case, we need only account for loadings and scores for a single factor; for models with $k > 1$, the sum of the products of loadings and scores over factors are needed.

The usual marginal predictions of family values across sites ($\widehat{Y}_{.j}$, the narrow-scope inference that includes the site-specific genetic deviations) are obtained as:



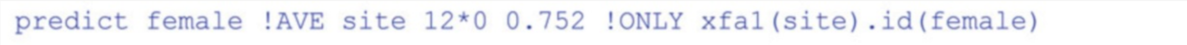
```
$B ~ mu site  !r  xfa1(site).id(female) ,
                  diag(site).id(block)
                  residual sat(site).id(units)
predict female  !PLOT female !AVE block site
```

Here we use `!AVE block site` to get the conditional predictions with appropriate standard errors for computing reliability (which we will show in the next section). The predictions for females across sites produce values similar to the other models (with differences due to allowing the within-site variances and between-site correlations to vary):

```
female          Predicted_Value  Standard_Error Ecode
  16            27.9630          0.1752 E
  18            26.4088          0.1815 E
...
 580            25.9551          0.2453 E
...
SED: Overall Standard Error of Difference   0.2439
```

The standard errors of the predictions are a bit smaller than in the previous model because of a better model fit. For female 580 at site 1, the SE of prediction is 0.2453 from XFA1 model compared to 0.2595 in CORUV model.

Prediction of family effects at a hypothetical ‘average’ environment can be accomplished with:



```
predict female !AVE site 12*0 0.752 !ONLY xfa1(site).id(female)
```

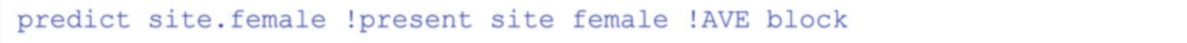
Here, the qualifier `!ONLY xfa1(site).id(female)` tells ASReml to make the prediction only using the parameter estimates of the XFA1 part of the model. The qualifier `!AVE site 12*0 0.752` refers to coefficients for the XFA1 model parameters: we set the coefficients for the first 12 parameters (the site-specific genetic variances) to zero to exclude them, and then we specify 0.752 as the average value of the site loadings on the first (and only) factor ($\bar{\lambda}_{1.}$).

The output from this predict statement in the *.pvs* file is an effect prediction, one could add it to the overall mean to get a predicted value:

```
female   Predicted_Value Standard_Error Ecode
   580   -0.4136         0.2583         E
```

The standard error of this predicted effect is a bit larger than the standard error for the average of site-specific predictions because it is predicted for a new, untested environment.

The predicted values of females at individual sites including the site-specific genetic deviation effects are easily obtained with:



```
predict site.female !present site female !AVE block
```

For example, the predicted value of family 580 at the first site is:

```
site   female   Predicted_Value   Standard_Error Ecode
 101    580     22.7646           0.6318         E
```

We can also obtain this predicted value as the sum of the intercept, the site 101 effect, and the predicted effect of family 580 at site 101:

$$\widehat{Y}_{1.580} = \mu + \widehat{S}_1 + \widehat{\lambda}_{1.1}\widehat{f}_{1.580} + \widehat{\delta}_{1.580}$$

The values needed to compute this prediction are found in the *.sln* file:

```
  Model_Term             Level     Effect    seEffect
 site                      101      0.000       0.000
...
 site                      113     -1.942      0.2195
 mu                          1      23.78      0.1989
diag(site).block       101.018    -0.1346      0.3062
...
xfa1(site).id(female   101.580     -1.016      0.6309
...
```

The term labelled ‘`xfa1(site).id(female`’ is the predicted genetic effect of family 580 at site 101, including the site specific genetic deviation:

$$\widehat{u}_{g1.580} = \widehat{\lambda}_{1.1}\widehat{f}_{1.580} + \widehat{\delta}_{1.580} = -1.016.$$

So the predicted value is: $\widehat{Y}_{1.580} = 23.78 + 0 - 1.016 = 22.764$, matching the prediction given directly in the *.pvs* file. We can also obtain the predicted effect of family 580 in site 101 based on only the FA1 part of the model as:

$$\widehat{u}^*_{g1.580} = \widehat{\lambda}_{1.1}\widehat{f}_{1.580}$$

We have already shown that the loading for site on the first factor is obtained in the *.asr* file:

```
Model_Term                Sigma    Sigma/SE  % C
...
site         XFA_L  1  1  0.824247     5.75  0 P
...
```

The factor score for family 580 is found in the last set of effect estimates in the *.sln* file. Note that the genotype factor scores are not printed out for the FACV or FA formulations of the model, only for XFA forms:

```
xfa1(site).id(female    1.580    -0.5501    0.3436
```

The predicted effect for this combination of family and site based only on the factor is:

$\widehat{u}^*_{1.580} = 0.824247^*(-0.5501) = -0.4533$, and the predicted value is:

$$\widehat{Y}^*_{1.580} = \mu + \widehat{S}_1 + \widehat{u}^*_{g1.580} = 23.78 + 0 + -0.4533 = 23.327$$

One can also obtain the factor-based family within site effects with a predict statement that excludes the site-specific genetic deviations:

```
predict female !AVE site 12*0 0.824 !ONLY xfa1(site).id(female)
```

This is very similar to the predict statement used previously to get the family effect prediction within a hypothetical ‘average’ environment, but in this case we use the loading for the first environment (0.824) instead of the average loading, resulting in the following prediction in the *.pvs* file:

```
female Predicted_Value Standard_Error Ecode
580    -0.4533         0.2831         E
```

Estimating Heritability and Reliability from FA Models

Estimating heritability as a function of observed variance components can be tricky when there are consolidated (compound) terms and complex covariance structures in the model, as in FA or US models. One difficulty is defining the appropriate function of variance components, for example if we have a model in which the genotypic variance is different for every environment. Understanding the labelling of parameter estimates in the function definitions in ASReml adds some additional complexity. Another difficulty can be having different mating designs such as half-sib families and full-sib families in the same data. In this case calculation of causal genetic variances (e.g. additive genetic variance) may not be obvious.

Before considering how to extend heritability estimates to complex MET models, it helps to consider the concepts of heritability, genetic variance, and environmental variance in the context of replicated family evaluation trials that often occur in tree and crop breeding experiments. Conceptually, the simplest assumption is that we have a reference population of genotypes from which the parents of the families are sampled, and, similarly, we have sampled the testing environments at random from the target population of environments, usually production environments within a defined geographic range (Cooper and DeLacy 1994). The variance components estimates for genotype main effects, environment main effects, and genotype-by-environment interaction effects refer to the variability in these conceptual reference populations (Dudley and Moll 1969).

In this context, the expected response to selection based on an individual's phenotype when its progenies are evaluated in an independent environment depends on the narrow-sense heritability, $h_i^2 = \sigma_A^2/\sigma_P^2$. We can estimate the pieces (additive genetic variance $\widehat{\sigma}_A^2$, and phenotypic variance $\widehat{\sigma}_P^2$) of this heritability estimator from a half-sib family evaluation like the pine polymix example using the traditional cross-classified analysis model as $\widehat{\sigma}_A^2 = 4\widehat{\sigma}_F^2$ and $\widehat{\sigma}_P^2 = \widehat{\sigma}_F^2 + \widehat{\sigma}_{FE}^2 + \widehat{\sigma}_\epsilon^2$. Thus, the narrow-sense heritability that is appropriate to predict response to selection among individual trees is:

$$h^2 = \frac{4\sigma_F^2}{\sigma_F^2 + \sigma_{FE}^2 + \sigma_\epsilon^2} \tag{8.26}$$

where σ_F^2 is the variance component due to family main effects, σ_{FE}^2 is the variance component due to family-by-environment interaction, and σ_ϵ^2 is the experimental error variance. Below, we will describe how to generalize this heritability estimator to more complex models such as the US and FA models with heterogeneous residual variances. Here, we will consider the appropriate heritability estimator to predict response to selection among family means. If we select superior families based on their means across environments and measure the response observed by growing remnant half-sib progenies in an independent environment sampled from the same reference population of environments, response to selection is a function of the selection differential and the heritability of family means defined using the cross-classified model structure as:

$$h_f^2 = \frac{\sigma_F^2}{\sigma_F^2 + \frac{\sigma_{FE}^2}{s} + \frac{\sigma_\epsilon^2}{sr}} \tag{8.27}$$

where s is the number of environments and r is the number of blocks per environment from which the means were calculated (Holland et al. 2003).

We can begin to generalize the estimator of family means-basis heritability by first considering the case where we have unbalanced data, with different numbers of plot measurements and environmental replications among families. One modification for unbalanced data is to use harmonic means of numbers of environments (s_h) and total plots (n_h) in which each family is measured (Holland et al. 2003):

$$h_f^2 = \frac{\sigma_F^2}{\sigma_F^2 + \frac{\sigma_{FE}^2}{s_h} + \frac{\sigma_\epsilon^2}{n_h}} \tag{8.28}$$

Another modification is the Cullis heritability estimator we introduced in Chap. 7 (Cullis et al. 2006):

$$h_{fC}^2 = 1 - \frac{\bar{V}_{BLUP\_difference}}{2\widehat{\sigma}_f^2} \tag{8.29}$$

The variance of the BLUP differences can be obtained from ASReml by squaring the average standard error of differences provided in the *.pvs* file when across-site family predictions are requested. Related to this estimator is the average of the prediction reliabilities, as introduced in Chap. 7.

A third modification is the bootstrapping method (Piepho and Möhring 2007). Note that no modification of the individual-basis narrow-sense heritability estimator is required when data are unbalanced because the selection units are individuals rather than family mean values.

To continue generalizing, when we have heterogeneous residual error variances, such that there are *s* distinct residual variances, the denominator of the narrow-sense heritability involves an average of the within-environment error variances:

$$h^2 = \frac{4\sigma_F^2}{\sigma_F^2 + \sigma_{FE}^2 + \overline{\sigma}_\epsilon^2} \tag{8.30}$$

Where $\overline{\sigma}_\epsilon^2$ is the average within-environment error variance. The *family mean-basis heritability estimate with heterogeneous error variances* includes a weighted average of within-environment variances:

$$h_f^2 = \frac{\sigma_F^2}{\sigma_F^2 + \frac{\sigma_{FE}^2}{s_h} + \frac{1}{s}\sum_{i=1}^{s}\frac{\sigma_{\epsilon i}^2}{r_{hi}}} \tag{8.31}$$

Where $\sigma_{\epsilon i}^2$ is the error variance within the *i*th environment and r_{hi} is the harmonic mean of number of plots per family in the *i*th environment. The Cullis estimator can also be used in this situation.

Finally, we generalize to the situation where the model has no genotype main effects, but rather *genotype effects nested in environments*. The response to selection among individual phenotypes as measured by their half-sib relatives grown in an independent environment is a function of a heritability estimator equal to the covariance of the selection and response individuals divided by the variance of individuals under selection (Nyquist 1991; Holland et al. 2003):

$$h^2 = \frac{E\left[Cov\left(f_{ij}, f_{i'j}\right)\right]}{V\left(f_{ij}\right)} \tag{8.32}$$

This is easily constructed from the estimated common genetic covariance between environments ($\widehat{\sigma}_{gii'}$) and common within-environment genetic variance component ($\widehat{\sigma}_{gi}^2$) from the CORUV model:

$$h^2 = \frac{\widehat{\sigma}_{gii'}}{\widehat{\sigma}_{gi}^2 + \overline{\sigma}_\epsilon^2} \tag{8.33}$$

For family-based selection, the predicted value of a family across sites is the average of its within-site predictions. To predict the response to selection based on this mean value as measured in an independent environment from the same reference population of environments used for the evaluation experiment, we want the expected covariance of the family mean to its value in an independent environment divided by the phenotypic variance of the family means:

$$h_f^2 = \frac{E\left[Cov\left(\bar{f}_{.j}, f_{i'j}\right)\right]}{V\left(\bar{f}_{.j}\right)} \tag{8.34}$$

Note that in the simple case of a model with family main effects and a common genotype-by-environment variance, the expected covariance of a family mean value with the family's value in an independent environment is estimated by the family variance component, and we have the usual heritability estimator for this model.

Considering the CORUV or compound symmetry model, we can use $\widehat{\sigma}_{gii'}$ and $\widehat{\sigma}^2_{gi}$ to estimate heritability:

$$E\left[\widehat{\sigma}_{gii'}\right] = E\left[Cov\left(\bar{f}_{.j}, f_{i'j}\right)\right] = \sigma_f^2$$

$$E\left[\widehat{\sigma}^2_{gi}\right] = E\left[V\left(ge_{ij}\right)\right] = \sigma_f^2 + \sigma_{fe}^2$$

$$V\left(\bar{f}_{.j}\right) = \sigma_f^2 + \frac{\sigma_{fe}^2}{s_h} + \frac{1}{s}\sum_{i=1}^{s}\frac{\sigma_{\epsilon i}^2}{r_{hi}} = \frac{(s_h - 1)\widehat{\sigma}_{gii'} + \widehat{\sigma}^2_{gi}}{s_h} + \frac{1}{s}\sum_{i=1}^{s}\frac{\sigma_{\epsilon i}^2}{r_{hi}}$$

$$\widehat{h}_f^2 = \frac{Cov\left(\bar{f}_{.j}, f_{i'j}\right)}{V\left(\bar{f}_{.j}\right)} = \frac{\widehat{\sigma}_{gii'}}{\frac{(s_h - 1)\widehat{\sigma}_{gii'} + \widehat{\sigma}^2_{gi}}{s_h} + \frac{1}{s}\sum_{i=1}^{s}\frac{\sigma_{\epsilon i}^2}{r_{hi}}} \tag{8.35}$$

If we have the more complex case of unequal pairwise variances among environments, our best estimate of the expected value of the covariance between the family mean value and its value in an independent environment is the average of the observed pairwise genotypic covariances between environments:

$$\widehat{Cov}\left(\bar{f}_{.j}, f_{i'j}\right) = \overline{Cov}\left(\bar{f}_{.j}, f_{i'j}\right) = \frac{1}{s(s-1)/2}\sum_{i=1}^{s-1}\sum_{i'=i+1}^{s}\widehat{\sigma}_{gii'} = \overline{\widehat{\sigma}_{gii'}} \tag{8.36}$$

The variance among family mean predictions is complicated if we have unbalanced data; it is the average over families of the variance of average family-by-environment effects:

$$\widehat{V}\left(\bar{f}_{.j}\right) = \bar{V}\left(\bar{f}_{.j}\right) = \frac{1}{n_f}\sum_{j=1}^{f} V\left(\frac{\sum_{i=1}^{S} ge_{ij}}{s_j}\right)^2 + \frac{1}{s^2}\sum_{i=1}^{s_j}\frac{\sigma_{\epsilon i}^2}{r_{hi}} \tag{8.37}$$

Here, the value s_j refers to the number of environments in which family j was tested. The effects of a common family at different environments are not independent, so we need to include the covariances among these terms as well as their variances in this case:

$$\begin{aligned}\bar{V}\left(\bar{f}_{.j}\right) &= \frac{1}{f}\left[\sum_{j=1}^{f}\frac{1}{s_j{}^2}\sum_{i=1}^{s_j} V\left(ge_{ij}\right) + \sum_{j=1}^{f}\frac{1}{s_j{}^2}\sum_{i=1}^{s_j}\sum_{i'\neq i}^{s_j} C\left(ge_{ij}, ge_{i'j}\right)\right] + \frac{1}{s^2}\sum_{i=1}^{s}\frac{\sigma_{\epsilon i}^2}{r_{hi}} \\ &= \frac{1}{f}\left[\sum_{j=1}^{f}\frac{1}{s_j{}^2}\sum_{i=1}^{s_j}\widehat{\sigma}^2_{gi} + \sum_{j=1}^{f}\frac{1}{s_j{}^2}\sum_{i=1}^{s_j}\sum_{i'\neq i}^{s_j}\widehat{\sigma}_{gii'}\right] + \frac{1}{s^2}\sum_{i=1}^{s}\frac{\sigma_{\epsilon i}^2}{r_{hi}}\end{aligned} \tag{8.38}$$

If data are balanced, this simplifies to:

$$\bar{V}\left(\bar{f}_{.j}\right) = \frac{\overline{\widehat{\sigma}^2_{gi}}}{s} + \frac{(s-1)\overline{\widehat{\sigma}_{gii'}}}{s} + \frac{1}{s^2}\sum_{i=1}^{s}\frac{\sigma_{\epsilon i}^2}{r} = \frac{\overline{\widehat{\sigma}^2_{gi}}}{s} + \frac{(s-1)\overline{\widehat{\sigma}_{gii'}}}{s} + \frac{\overline{\sigma_{\epsilon i}^2}}{sr} \tag{8.39}$$

Translating this to the model with family main effects, the variance of family mean values is:

$$\begin{aligned}\bar{V}\left(\bar{f}_{.j}\right) &= \frac{\sigma_F^2 + \sigma_{FE}^2}{s} + \frac{(s-1)\sigma_F^2}{s} + \frac{\overline{\sigma_{\epsilon i}^2}}{sr} \\ &= \sigma_F^2 + \frac{\sigma_{FE}^2}{s} + \frac{\overline{\sigma_{\epsilon i}^2}}{sr}\end{aligned} \tag{8.40}$$

This simplifies further in the case of homogenous error variances to the standard estimator of heritability from multi-environment trials with balanced data:

$$\bar{V}\left(\bar{f}_{.j}\right) = \sigma_F^2 + \frac{\sigma_{FE}^2}{s} + \frac{\sigma_\epsilon^2}{sr} \tag{8.41}$$

Putting the average covariance between families across environments as the numerator and the average variance of family means across environments as the denominator as the heritability estimate, we get for the case of unbalanced data and heterogeneous genetic variances and covariances across sites:

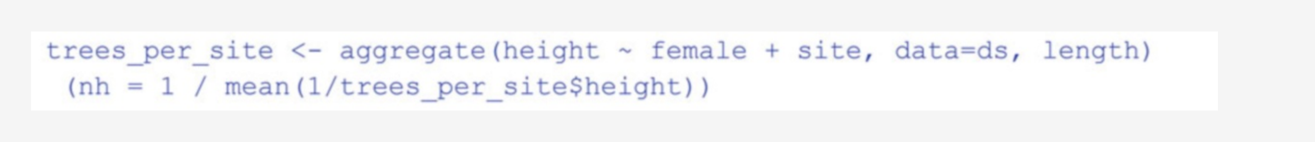
$$h_f^2 = \frac{\overline{\widehat{\sigma}_{gii'}}}{\frac{1}{f}\left[\sum_{j=1}^{f}\frac{1}{{s_j}^2}\sum_{i=1}^{s_j}\widehat{\sigma}_{gi}^2 + \sum_{j=1}^{f}\frac{1}{{s_j}^2}\sum_{i=1}^{s_j}\ \sum_{i'\neq i}^{s_j}\widehat{\sigma}_{gii'}\right] + \frac{1}{s^2}\sum_{i=1}^{s}\frac{\sigma_{\epsilon i}^2}{r_{hi}}} \tag{8.42}$$

In the case of balanced data but heterogeneous error variances, this simplifies to:

$$h_f^2 = \frac{\overline{\widehat{\sigma}_{gii'}}}{\frac{\overline{\widehat{\sigma}_{gi}^2}}{s} + \frac{(s-1)\overline{\widehat{\sigma}_{gii'}}}{s} + \frac{\overline{\sigma_{\epsilon i}^2}}{sr}} \tag{8.43}$$

Now we will use the parameter estimates from different models for the pine polymix data to estimate heritability of family means across environments. About 3% of plots are missing in this data set, so we should use Eq. 8.42, which involves the mean within-site variances weighted by the harmonic mean of replications per family at each site, but for simplicity, and because the level of imbalance is low, we will use the balanced data formula (Eq. 8.43), substituting the harmonic mean of the number replications per family and site (17.8) for the value *r*.

The harmonic mean of trees per family per site can be computed easily in R from a data frame (called “ds” in this example) holding our data:



```
trees_per_site <- aggregate(height ~ female + site, data=ds, length)
 (nh = 1 / mean(1/trees_per_site$height))
```

First, we estimate narrow-sense and family mean-basis heritabilities for the *cross-classified* MET model with homogeneous error variances:

$$h^2 = \frac{4\sigma_F^2}{\sigma_F^2 + \sigma_{FE}^2 + \sigma_\varepsilon^2} = \frac{4(0.563)}{0.563 + 0.174 + 7.12} = 0.29$$

$$h_f^2 = \frac{\sigma_F^2}{\sigma_F^2 + \frac{\sigma_{FE}^2}{s} + \frac{\sigma_\varepsilon^2}{rs}} = \frac{0.563}{0.563 + \frac{0.174}{12} + \frac{7.12}{17.8*12}} = 0.92$$

Using the parameter estimates (rounded to the third decimal) from the compound symmetry (CORUV) model, we get the same results:

$$h^2 = \frac{4r_g\widehat{\sigma}_{gii'}}{\widehat{\sigma}_{gi}^2 + \sigma_\varepsilon^2} = \frac{4(0.763)(0.738)}{0.738 + 7.12} = 0.29$$

$$h_f^2 = \frac{\overline{\widehat{\sigma}_{gii'}}}{\frac{\overline{\widehat{\sigma}_{gi}^2}}{s} + \frac{(s-1)\overline{\widehat{\sigma}_{gii'}}}{s} + \frac{\sigma_\varepsilon^2}{rs}} = \frac{r_g\widehat{\sigma}_{gi}^2}{\frac{\overline{\widehat{\sigma}_{gi}^2}}{s} + \frac{(s-1)r_g\widehat{\sigma}_{gi}^2}{s} + \frac{\sigma_\varepsilon^2}{rs}}$$

$$= \frac{(0.763)(0.738)}{\frac{0.738}{12} + \frac{11(0.763)(0.738)}{12} + \frac{7.12}{17.8*12}} = 0.92$$

The estimates for the CORUV model with heterogeneous error variances are:

$$h^2 = \frac{4r_g\widehat{\sigma}_{gii'}}{\widehat{\sigma}^2_{gi} + \overline{\sigma^2_{\varepsilon i}}} = \frac{4(0.8428)(0.6607)}{0.6607 + 7.21} = 0.28$$

$$h^2_f = \frac{r_g\widehat{\sigma}^2_{gi}}{\frac{\overline{\widehat{\sigma}^2_{gi}}}{s} + \frac{(s-1)r_g\widehat{\sigma}^2_{gi}}{s} + \frac{1}{s^2}\sum_{i=1}^{s}\frac{\sigma^2_{\varepsilon i}}{r}}$$

$$= \frac{(0.843)(0.661)}{\frac{0.661}{12} + \frac{11(0.843)(0.661)}{12} + \frac{10.23 + 3.18 + \ldots + 9.40}{12^{2*}17.8}} = \frac{0.557}{0.598} = 0.93$$

Finally, for the XFA1 model with heterogeneous error variances, recall that the lower diagonal of the variance-covariance matrix of family within environment effects is (for 4 sites out of 12):

```
1.449 0.6288 0.6445 0.6848....

0.6260 0.6848 0.6445 0.6755....

0.5595 0.5464 0.8643 0.9183....

0.8395 0.9183 0.8643 0.9059....

...
```

The heritability estimate is based on the estimated variance-covariance matrix of family within environment effects. For this model, the average of the diagonal elements (0.7408) is the mean within-site family variance, and the average of the off-diagonal elements (0.563) is the average covariance between sites. The mean of the 12 site-specific residual variances is 7.15:

$$h^2 = \frac{4\overline{\widehat{\sigma}_{gii'}}}{\widehat{\sigma}^2_{gi} + \overline{\sigma^2_{\varepsilon i}}} = \frac{4(0.563)}{0.741 + 7.15} = 0.28$$

In the heritability for individual measurements, the mean within-site family variance in the numerator is multiplied by 4 because the mean within-site female variance is 1/4 of the additive genetic variance due to the half-sib family structure in the data. In contrast, for the estimate of heritability on a family mean basis, we use the family variance component directly in the numerator, since our inference is to selection among the family mean predictions:

$$h^2_f = \frac{\overline{\widehat{\sigma}_{gii'}}}{\frac{\overline{\widehat{\sigma}^2_{gi}}}{s} + \frac{(s-1)\overline{\widehat{\sigma}_{gii'}}}{s} + \frac{1}{s^2}\sum_{i=1}^{s}\frac{\sigma^2_{\epsilon i}}{r_{hi}}} = \frac{0.563}{\frac{0.741}{12} + \frac{11(0.563)}{12} + \frac{7.15}{12^{2*}17.8}} = 0.92$$

Again, for the XFA1 model we obtained similar narrow-sense and family mean heritabilities.

We can also estimate the family mean-basis heritability using the Cullis estimator by taking the average of the standard error of across-site family differences (SED) from the *.pvs* file:

```
SED: Overall Standard Error of Difference 0.2438
```

$$h^2_{fC} = 1 - \frac{\bar{V}_{BLUP\_difference}}{2\widehat{\sigma}^2_f} = 1 - \frac{(0.2438)^2}{2\left(\overline{\widehat{\sigma}_{gii'}}\right)} = 1 - \frac{(0.2438)^2}{2(0.563)} = 0.947$$

Table 8.2 Family predictions across sites from model 11 (XFA1) and their reliabilities

Female	Predicted_Value	Standard_Error	REL
16	27.9630	0.1638	0.952
18	26.4088	0.1705	0.948
414	27.5957	0.1711	0.948
...			
580	25.9551	0.2372	0.900
...			
Average			0.947

This is close to the family mean-basis heritability based on variance components.

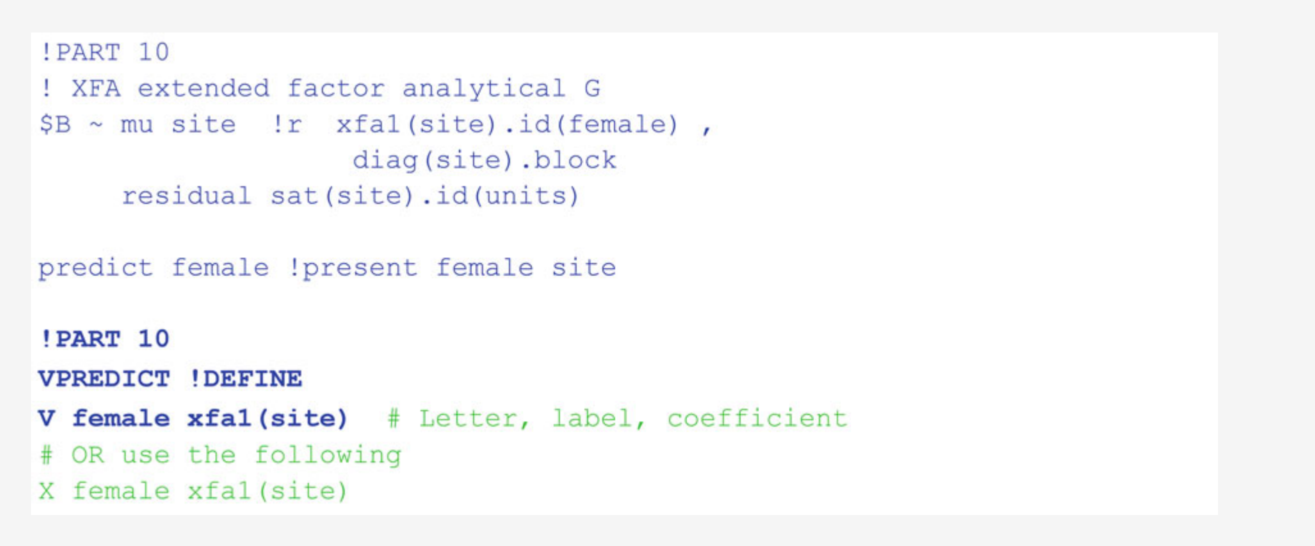
One more way to estimate the family mean-basis heritability is as the average of the prediction reliabilities, using the formula:

$$REL = 1 - \frac{PEV}{\sqrt{\sigma_f^2}} = 1 - \frac{PEV}{\sqrt{\overline{\sigma}_{ii'}}} \tag{8.44}$$

From the XFA model, we will use 0.563 in the denominator of the reliability equation. From the *.pvs* output of the statement ‘`predict female !AVE block site`’ in model 11, we can compute reliabilities (Table 8.2):

The average of the reliabilities (0.947) is identical to the Cullis estimator of family mean heritability.

We can obtain the estimates based on functions of variance components using the VPREDICT !DEFINE option in ASReml. As shown in Chap. 6, the easiest way to get the correct labels of parameter estimates from a complex ASReml model is to use VPREDICT !DEFINE at the end of the model and leave a blank line after it to generate a *.pvc* file with names and numbers of parameters identified. In this example we will estimate heritability from the XFA1 structure in model 11.



```
!PART 10
! XFA extended factor analytical G
$B ~ mu site  !r  xfa1(site).id(female) ,
                   diag(site).block
     residual sat(site).id(units)

predict female !present female site

!PART 10
VPREDICT !DEFINE
V female xfa1(site)   # Letter, label, coefficient
# OR use the following
X female xfa1(site)
```

- The components labeled ‘female’ were created using **`V female xfa1(site)`**.
- **`V`** is the function to convert components from XFA to unstructured (US) model parameters (i.e., to provide the within-environment variances and each of the pairwise environment covariances), ‘female’ is the label we assign and ‘`xfa1(site)`’ is the identifier of the variance component.

The output found in “Code 8-1_MET11_height.pvc” is given below:

```
          - - - Results from analysis of height - - -

sat(site,01).id(units)        1359 effects
   1 sat(site,01).id(units);Residual_1           9.81099   0.389325
...
  12 sat(site,12).id(units);Residual_12          5.46531   0.218263
...
xfa1(site).id(female)          910 effects
  25 xfa1(site).id(female);xfa1(site)  V  0  1  0.769546      0.226337
...
  36 xfa1(site).id(female);xfa1(site)  V  0 12  0.126178E-01  0.525742E-01
  37 xfa1(site).id(female);xfa1(site)  L  1  1  0.824246      0.149050
...
  48 xfa1(site).id(female);xfa1(site)  L  1 12  0.675730      0.803484E-01
  49 female      1.4489   0.32321     (variance site 1)
  50 female     0.55949   0.13227     (cov 1,2)
  51 female     0.54638   0.11257      (variance site 2)
  52 female     0.59378   0.14588      (cov 1,3)
  53 female     0.48899   0.97253E-01 (cov 2,3)
  54 female     0.58588   0.15189     (variance site 3)
  55 female     0.61280   0.14262     (cov 1,4)
...
 124 female     0.50422   0.97664E-01 (cov_9,12)
 125 female     0.66804   0.13021     (cov_10,12)
 126 female     0.46923   0.11384     (cov_11,12)

Notice: The parameter estimates are followed by
          their approximate standard errors.
```

- Coefficients are identified by the numbers in the first field and by labels. For example, residual variances for sites are numbered from 1 to 12, and labeled as

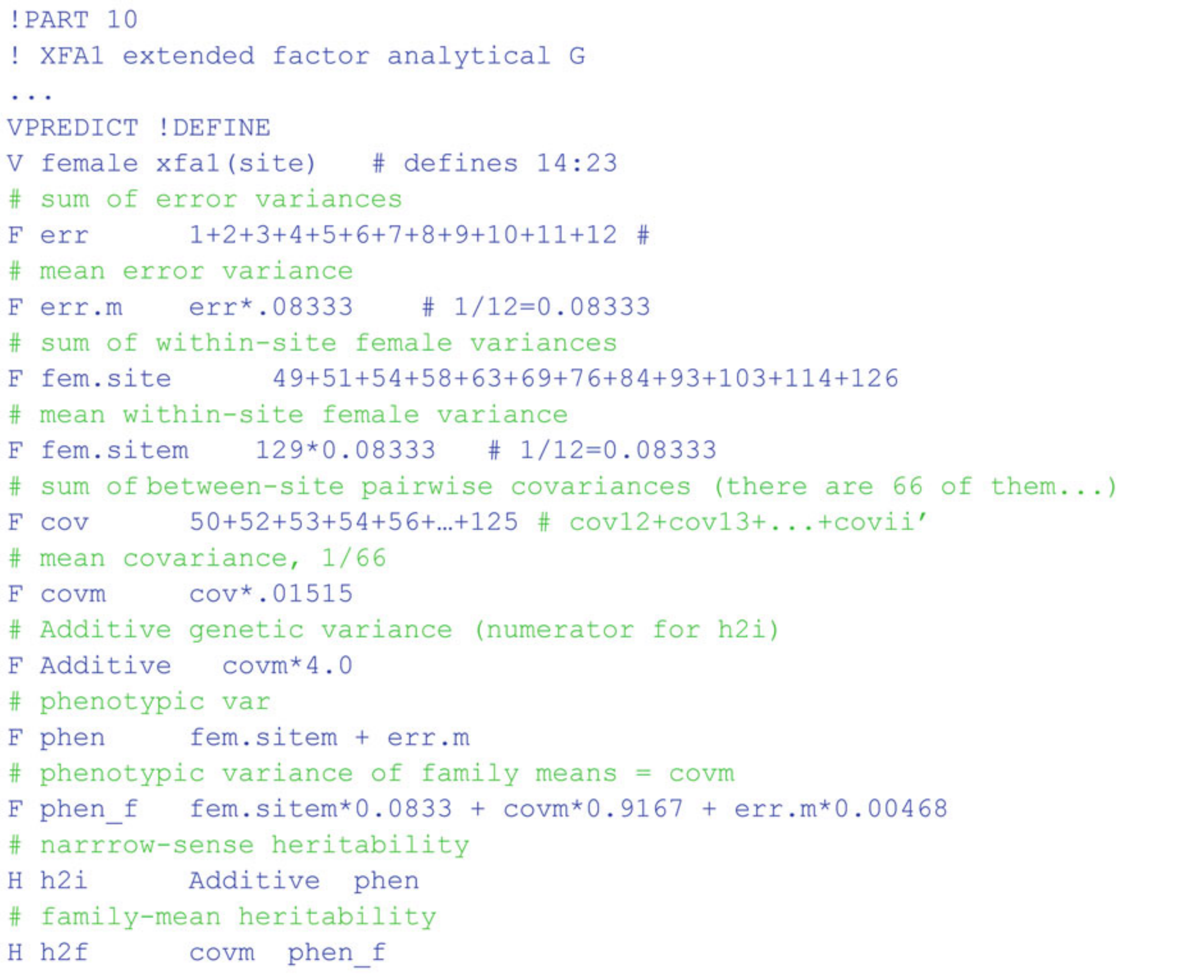
```
1 sat(site,01).id(units);Residual_1
2 sat(site,02).id(units);Residual_2
```

- The fields named `female` (numbered from 49 to 126) are female within-site variance components (bold) and covariances between pairs of site. If we rearrange them in matrix format it will be more obvious how they relate to the US parameterization (for the first 4 sites):

```
       site1    site2   site3   site4
site1  1.449
site2  0.5595   0.5464
site3  0.5938   0.4890  0.5859
site4  0.6128   0.5047  0.5356  0.5527
```

In the following example, we compute phenotypic variances, additive genetic variances, and heritabilities for selection among individual trees or family means.

PART 10



```
!PART 10
! XFA1 extended factor analytical G
...
VPREDICT !DEFINE
V female xfa1(site)   # defines 14:23
# sum of error variances
F err      1+2+3+4+5+6+7+8+9+10+11+12 #
# mean error variance
F err.m    err*.08333    # 1/12=0.08333
# sum of within-site female variances
F fem.site      49+51+54+58+63+69+76+84+93+103+114+126
# mean within-site female variance
F fem.sitem    129*0.08333   # 1/12=0.08333
# sum of between-site pairwise covariances (there are 66 of them...)
F cov      50+52+53+54+56+…+125 # cov12+cov13+...+covii’
# mean covariance, 1/66
F covm     cov*.01515
# Additive genetic variance (numerator for h2i)
F Additive   covm*4.0
# phenotypic var
F phen     fem.sitem + err.m
# phenotypic variance of family means = covm
F phen_f   fem.sitem*0.0833 + covm*0.9167 + err.m*0.00468
# narrrow-sense heritability
H h2i      Additive  phen
# family-mean heritability
H h2f      covm  phen_f
```

A subset of the output (*Code 8-1_MET11_height.pvc*) is given below:

```
         - - - Results from analysis of height - - -
...
 127 err  1            85.820        1.0644
 128 err.m127          7.1514       0.88693E-01
 129 fem.site 49       8.8899        1.1220
 130 fem.sitem129     0.74079       0.93495E-01
 131 cov 50           37.190         5.6923
 132 covm131         0.56348       0.86247E-01
 133 Additive132       2.2539       0.34499
 134 phen128           7.8922       0.12610
 135 phen_f130        0.58104       0.86409E-01
     h2i          = Additive132 133/phen128  134=  0.2856 0.0407
     h2f          = covm131  132/phen_f13 135=  0.9211 0.0105
 Notice: The parameter estimates are followed by
          their approximate standard errors.
```

These estimates agree with our computations above. As the number of environments increases, the number of covariance for pairs of environment increases (66 in this example). This makes heritability calculations cumbersome in ASReml. Care is needed to make sure variances and covariances are selected correctly.

Biplots from FA Models

Biplots can be useful visualizations of GxE interactions from FA models. The responses of genotypes to environments on a two-dimensional surface are frequently reported in the plant breeding literature. A biplot displays site loadings and genotype scores simultaneously. R code to read in results from an XFA2 model produced by ASReml standalone and to generate a biplot is provided in "*Code 8-3_biplot.R*". Another form of the biplot using output of ASReml-R is provided in "*Code 8-2_pine_met.Rmd*". Figure 8.1 was produced by the first set of code and displays site loadings as vectors in blue on the two factors and family scores in black. This figure shows a typical problem with biplots: if the number of genotypes or families is large, the plot becomes very busy and hard to read. Nevertheless, even from this plot, it is clear that site 105 affected genotype performance differently than other sites. This is congruent with the result observed in the correlation estimates from the XFA models that indicate that this site had the lowest average correlation with other sites. A large number of genotypes are at the center of the plot; genotypes that are closer to the end of a particular site vector have scores with the same sign and similar magnitude of that site's loading compared to the rest of the population. This indicates that those families have their most positive effect at that environment. For example, family 421's score is near the loading for site 105, indicating that it has the most favorable effect at that site. Indeed, family 421 has the highest predicted value at site 105 (30.7, compared to a population mean of 28.9 at that site). Biplots are descriptive, however, and should be interpreted cautiously as they may not depict all aspects of the GxE interactions, including crossover interactions (Yang et al. 2009).

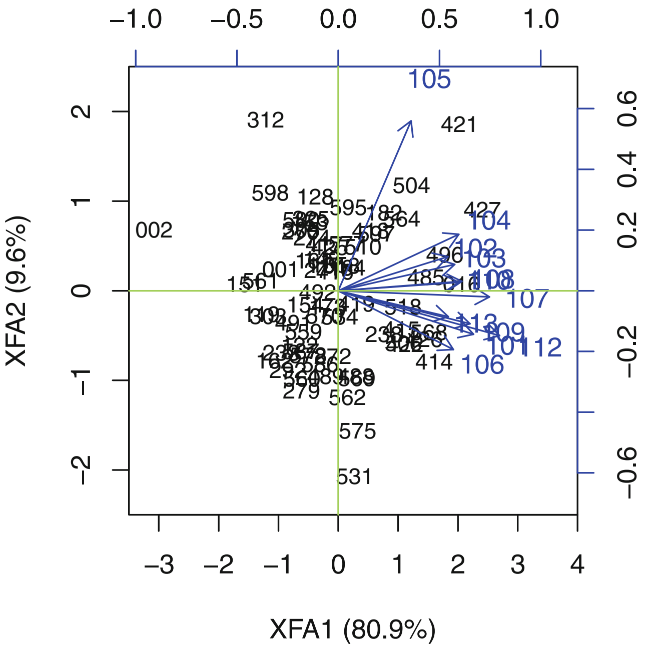


Fig. 8.1 Biplot for site loadings (*blue vectors*) and female scores (*black text labels*) on two factors from the XFA2 model. Site 105 has high within-site variation but has the smallest correlation with other sites. Genotypes 421, 504 and 427 have large positive scores for both factors